Appraising Research in Second Language Learning: A practical approach to critical analysis of quantitative research

Graeme Keith Porte

John Benjamins Publishing Company

Appraising Research in Second Language Learning

Language Learning and Language Teaching

The *LL&LT* monograph series publishes monographs as well as edited volumes on applied and methodological issues in the field of language pedagogy. The focus of the series is on subjects such as classroom discourse and interaction; language diversity in educational settings; bilingual education; language testing and language assessment; teaching methods and teaching performance; learning trajectories in second language acquisition; and written language learning in educational settings.

Volume 3

Appraising Research in Second Language Learning: A practical approach to critical analysis of quantitative research
by Graeme Keith Porte

Appraising Research in Second Language Learning

A practical approach to critical analysis of quantitative research

Graeme Keith Porte
University of Granada

John Benjamins Publishing Company
Amsterdam / Philadelphia

Dr Graeme Keith Porte is currently Senior Lecturer at the University of Granada in Spain, where he lectures on Second Language Writing and Applied Linguistics Research Design. He frequently evaluates papers for international publications in the field and currently serves as an editorial advisor to the board of "LANGUAGE TEACHING" (Cambridge University Press)

Library of Congress Cataloging-in-Publication Data

Porte, Graeme Keith
Appraising research in second language learning : a practical approach to critical analysis of quantitative research / Graeme Keith Porte.
p. cm. (Language Learning and Language Teaching, ISSN 1569-9471 ; v. 3)
Includes bibliographical references and index.
1. Second language acquisition--Research--Methodology. I. Title. II. Series.

P118.2 P66 2002
418'.007'2-dc21 2002074683
ISBN 90 272 1695 9 (Eur.) / 1 58811 253 5 (US) (Hb; alk. paper)
ISBN 90 272 1696 7 (Eur.) / 1 58811 254 3 (US) (Pb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

To my father, who encouraged me to ask questions

Table of contents

Preface

This book is written to guide student and novice researchers through their critical reading of a research paper in the field of second language learning. My aim is to help these readers relate the basic knowledge they acquire during introductory courses on investigation in applied linguistics to their own independent reading of research papers. They will be shown ways of approaching the appraisal of the abstract and the introductory section of the study, both of which set the stage by describing the rationale as well as the objective of the work. Similarly, the reader will be given ideas about how to assess the method and procedures section so that he or she can decide, for example, whether the research design was appropriate, and what precautions were taken to guard against threats of validity to the findings. They will become more familiar with, and confident about, interpreting results from commonly-used descriptive or inferential statistical procedures and checking how appropriately these have been presented. Finally, the reader should be in a position critically to evaluate the researcher's own interpretation of the findings in terms of the extent to which the conclusion is justified, can be generalised, and has limitations.

An experienced and critical reader will also contribute enormously to research practice. Above all, there is the obvious help and experience critical reading gives us towards the better description and presentation of our own studies. Informed criticism of others' work inevitably also helps us to discover new areas of research that have emerged as a direct result of this critical reading. This said, I will *not* be focussing here specifically on designing and conducting a critique of research as direct preparation for carrying out, writing up, and publishing one's own study. Indeed, as will become clear in the text, different journals and other media adopt different policies with regard to the presentation and writing up of research for publication, varying in accordance with the orientation and objectives of the particular medium. Nevertheless, learning to read research appropriately is, indeed, intimately connected with learning to write it effectively, and it is to be hoped that the experience of appraising in this

way will help the reader better to present their own work for publication and peer evaluation.

The emphasis in the practical application sections of the book will be on so-called "quasi-experimental studies", using participants in L2[1] classroom situations. I have adopted this stance because such designs are more representative of the conditions typically found for research in educational contexts and, specifically, because so much of the research currently undertaken and published in our area involves the use of classes to which subjects have already been assigned following certain internal guidelines. Such "intact groups" imposed by the local administration often mean that it is difficult — on occasions impossible — satisfactorily to meet the many threats to data validity present in such research, in particular those relating to selection of subjects, allocation of groups, and experimental procedures. Conversely, for many of us working in the area of second language learning, undertaking studies with intact groups is the only practical way of conducting research. This is not to belittle the contribution already made — and still to be made — by such studies: our research objectives will always need to take such unavoidable constraints into account. But this reality means that we should learn how to use such familiar designs to our advantage and, to the best of our ability, search out answers to our research hypotheses and questions, while acknowledging the inevitable limits imposed on the interpretation of results obtained in these research contexts.

Why do we need to appraise research?

The importance of the book lies in the fact that it responds to a current need in the field. Students and potential researchers may have read many academic papers and absorbed considerable theory about research design and how to implement it in their own study. Indeed, more and more universities and teaching colleges are including courses in their degrees aimed at providing students with a basic introductory knowledge of research techniques and practice. This is a particularly timely response to a perceived need for students who will ideally want to contribute to their chosen field of study. However, conducting research is much more than merely knowing how to carry out the investigation.

---

1. Except where indicated, "L2" refers to both second language and foreign language learning contexts.

Sound preparation for research also requires the ability to situate and defend our proposed study in the light of existing knowledge. In order to do this, we need to be able to appraise others' work adequately and appropriately enough to help justify the contribution our own work is intended to make to current awareness in the field.

Although my assumption is that all of us work and/or study in the area of second language learning, our interest in academic papers will inevitably vary according to our own personal line of research. Thus, while the extracts and sample papers the reader is going to study in this book reflect important areas of research, we will inevitably all be approaching a paper from different backgrounds, interests, and needs. Nevertheless, our overall critical approach to the object of our reading will be similar. It will, therefore, always be important for anyone involved in any kind of research in our area to be able to approach the reading of a study both from the point of view of an uncommitted critical reader and/or that of a researcher interested in one focal aspect of that area.

Who is this book for?

This book is primarily intended as a main course text (or supplementary to a "research-techniques" book such as those mentioned below) and is aimed at classes of both undergraduate and postgraduate students reading for language or applied linguistics degrees who are required to submit work which entails the understanding of the theory and practice of research principally based on quantitative analyses. It also links with the growing number of "applied" professional courses and should prove of considerable interest to those participating in language teacher-training programmes or degrees, as well as those practising teachers anxious to embark on their own classroom research. Both groups need to be "research literate" and, as Altman (1988) maintains, "second language teachers have a professional obligation to make sense of research that has a potential impact on their classrooms"[2] — even if they do not wish actively to participate in their own research.

Appraisal of research papers not only requires common sense, but also some degree of literacy in this kind of research. To take full advantage of this

2. Foreword to Brown, J. 1991. *Understanding Research in Second Language Learning.* Cambridge: Cambridge University Press, p vii.

text, the reader should already understand the basic principles of research in this field and needs to have some rudimentary knowledge of how the most common inferential and descriptive statistical procedures — as in correlation or comparing means and frequencies — might be used in data analyses. A small number of excellent books are already available which provide an introduction to terminology and the most useful quantitative and qualitative techniques typically used in research in the area of second language teaching and learning (see Appendix, "Further Reading"). Where descriptive or inferential statistical procedures *are* discussed in this text, the emphasis is on the appropriateness and consequences of their application in the context of what we have read in the paper, rather than on the way these procedures should actually be carried out. Furthermore, a large number of illustrative examples and a "Glossary of key terms in quantitative research" have been included to facilitate the critical process. Nevertheless, the book provides only an introduction to key issues in certain statistical matters, such as the advantages and disadvantages of parametric and non-parametric procedures, or the implicit assumptions to be applied when using a particular test, and always strictly in terms of the way such concerns have an impact on the way we appraise any outcomes. Readers who intend to move beyond critique to employing some of these procedures in their own research need to obtain more profound knowledge and are referred to more detailed accounts in the recommended books.

How is the material to be used?

An important objective of this book is to provide the reader with a suggested methodology that can eventually be applied to his or her own independent appraisal of quantitative research. The appraisal of a research paper requires the ability to be attentive throughout the reading and thereby be able to react to the text based both on what we know and what we have been told. Continually stimulating and channelling such reactions to the reading forms the basis of the learning approach applied in this book. The principal innovative methodology used here to arouse the reader's critical faculties is awareness-raising. The reader is encouraged to react to information at the moment of reading rather than — as usually happens — in a subsequent review of the whole text. Thus, my assumption throughout is that the critical reader approaches their reading in the usual sequential manner. In other words, they will normally evaluate each section as and when they meet it, without looking ahead to other sections of the

paper, but recalling what they have already read. I have adopted such a progressive *modus operandi* not only out of practical considerations, but also because it represents a very sound reading strategy when one is learning how to appraise research: it encourages the critical reader to begin to build up a response to what they are reading. Notwithstanding these assumptions, there will be times within the activities when a reader is encouraged to look ahead to what is said at later stages in the paper for further possible insights.

What this means in practice within an awareness-raising methodology is that, at various points in the text, I encourage readers to pause in their reading, to summarise and question what has just been read, to ponder over the consequences of a statement, to predict outcomes, to think back to previous parts of the paper and ahead to upcoming sections, or to suggest possible problems or drawbacks in the perceived research design. Obviously, such spontaneous responses are only expressions of our preliminary reactions to something, and they may well be refuted by what subsequent sections of the paper reveal. Consequently, they may or may not need to be modified in the light of that information. However, the exercise of raising our awareness and reacting to the text instinctively in this way is particularly helpful for a reader who is seeking to engage critically with the text. Essentially, it draws us immediately into the research itself and puts us in the role of an inquisitive observer, almost as if we were engaging in a real-time conversation with the author about the study.

Here a word of warning. A popular notion exists that appraising a fellow-researcher's work consists of finding faults in that work. Given what should be the fundamentally supportive role of the critical reader of research, I would like to dissuade you from this notion at once. Appraisal of scientific research can be approached in many different ways but, by definition, will require us to adopt a cautious attitude to what we read and will involve us making judgements about the perceived quality and merit of what has been described. However, such remarks should be seen to have a constructive — rather than *de*constructive — aim both for the author *and* the reader. As scientists and readers of such science it is in the interest of our field that we are sceptical; it is also by asking pertinent questions of the research we read that we ourselves may better shape the course of our *own* future research. We should not expect, as researchers or readers of research, to find a study which provides us with *all* the answers to *all* our questions. However, our reading should at least help put us in a better position to decide whether all the appropriate questions were being asked in the first place. Such appraisal needs to respond appropriately to the most consequential aspects of the study, rather than to some kind of

"check-list" of essential elements of any research paper. Indeed, Gottfredson (1978)[3] identified no less than 83 attributes that many editors/reviewers may use in their appraisals. Given such a large number of elements, it is not surprising that many journal reviewers themselves end up in disagreement about the merits of a paper!

The main idea behind the awareness-raising approach to appraisal encouraged here is that readers see themselves not as judge and jury of a study, but rather as potential consumers of research. As such, we will initially need to evaluate the contribution and importance of the work to our own present interests. In turn, learning how to appraise will enable us to assess the amount of confidence we might reasonably have both in the findings and the interpretations made from these. A well-designed study will provide answers in which we can have confidence and will serve as an example, and perhaps a stimulus, for our own work. In this sense, the overall significance of this book is that it shows the reader how to learn directly and indirectly from what he or she is reading.

This book has been written in a combination text- and workbook format. In the **textbook** section, the reader is introduced to the most typical component parts of a quantitative research paper in this field. Here a personal perspective on the paper is provided which explains and illustrates ways in which essential information may ideally be communicated to the reader and critically interpreted within that part. There is inevitably a degree of subjectivity involved in one reader appraising certain aspects of L2 research; however, I have tried throughout to provide authoritative advice on appraisal, rather than impose an authoritarian approach.

For the purposes of this approach to appraisal, a research paper is divided into four basic elements (i. introduction; ii. method and procedures; iii. results, and iv. discussion and conclusions) and a number of subsections within these four. In the "INTRODUCTION" chapter readers will be presented with strategies for appraising the abstract, the background to the problem and the problem statement, the review of the literature, research objectives and variables, and operational definitions. Obviously, an adequate understanding of the value of the whole paper cannot be achieved without appreciating the sum of its parts. The reader is shown how the abstract may provide both context and landmarks: the kind of information which enables the reader to assess the

3. Gottfredson, S. 1978. Evaluating psychological research reports. *American Psychologist,* 33, 920–934.

immediate value the paper has for his or her current interests. Subsequently, the reader is taken through the process of locating and evaluating the background to the study and the problem statement and interpreting any variations found during the paper. The reader then decides how far the literature cited provides an adequate theoretical and empirical basis for the subsequent development of the study's hypotheses and/or research questions. Finally, he or she is encouraged to think about the operational descriptions and assignment of variables and constructs in the study. The advice and practice given in this section is aimed at helping the reader be in a better position to handle a posterior evaluation of the "Results" and "Discussion" sections of the paper being read, both in the light of the information provided here and the responses made to it.

Once it is established *what* the text tells us about the study and the background to, and nature of, the research question or hypothesis, we might look for a suitable description of *how* the study was carried out. Again, the recommended appraisal process here is one of constant action on the part of the author and reaction on the part of the reader. Thus, what is explained to us in this part of a paper will need continually to be appraised in the light of our reactions to the information in the "Introduction". In the METHOD AND PROCEDURES chapter the following areas are treated: Internal and external validity issues; subject identification, processes of selection and group assignment; procedures; principal research designs; and proposed measurement and analysis. Thus, this chapter deals with the nuts-and-bolts of the research design. Issues which need to be appraised here include how subjects have been selected for study and the consequences of such selection for any subsequent interpretation, the conditions in which subjects are observed and the limitations imposed by such circumstances, the instructions given to the subjects and the implications of these for the data obtained, or the way in which research designs with intact groups may differ from true experimental designs. Identifying a study's research design is helpful when evaluating the procedures described by the author. Since research design is not normally described by the authors themselves, it will often be up to the reader to assess the suitability of the procedures used. To this end, the reader is shown the benefits of considering designs graphically, highlighting their advantages and disadvantages using the help given. In this way, the reader is also encouraged to comment on any potential weaknesses in the design and suggest possible improvements.

Throughout, particular attention is paid to precision in replicability of designs. If the descriptions we read are vague or do not allow for replication, then in some way they are not precise enough to allow the reader to evaluate

and interpret the information given in the subsequent "Results" section. Similarly, another of the interesting questions considered here is the appraisal of any proposed data analyses. After an introduction to the most commonly-used statistical analyses of data, the reader is encouraged to appraise with a view to verifying the extent to which the researcher has thought about and checked the assumptions underlying any proposed statistical analyses and/or whether he or she has allowed for the potential effects of violations on subsequent results.

In the RESULTS chapter, the focus shifts to outcomes and their initial interpretation. In many cases, in the kind of study I am concerned with here, the selection of an appropriate analytical procedure will be extremely important to establish confidence in results — both for the author and the reader. That confidence, hopefully gained in previous sections of the paper, will ideally now be reinforced through the nature and presentation of the findings. Readers will need to see how to locate the most important results (i.e., those which have a direct bearing on the proposed problem statement). Consideration for appraisal will include whether these have been stated or presented clearly enough for the reader to interpret them appropriately. To this end, the reader is shown how to make use of any tables or figures provided in the paper, particularly those that provide the basis for checks to be made on claims of relationships, similarities, or differences between variables. The reader is subsequently encouraged to assess the results as a consequence of the research context, the objectives of the study, the subjects used, and the choice of instrument used for analysis. Therefore, particular attention is paid here to helping the reader assess what they are being told based on the information given and appraised earlier in the paper. The reader is also encouraged to check on typical but questionable practices, such as doing additional tests when the planned ones did not provide significant results, or using parametric statistical tests without enough evidence of normality in the original database.

Within this section, both statistical and practical significance (or meaningfulness) of results will also be a source of concern. Specifically, readers are shown how they might go about evaluating the significance of findings through any text or tables provided and, if these are reported, to see whether such outcomes actually represent the major question asked or some subsidiary, post-hoc question not really part of the original problem statement. As regards meaningfulness, the aim is to assess whether any apparently significant data reported are, in fact, interesting or important once placed within the context and objectives of the study.

In the final chapter, strategies are suggested to assess the quality of the DISCUSSION AND CONCLUSIONS. The first focus of appraisal in this section is the extent to which any conclusions drawn are consistent with the results. The more studied interpretations we might expect to encounter in this section of the paper will need to be evaluated in terms of their reasonableness, and any explanations for unexpected outcomes carefully considered. If, for example, an author claims that a particular variable did not affect results in the way predicted, readers may want to judge how far the reason given is acceptable in the circumstances. Similarly, questions of generalisation and practical implications of findings are appraised here. The reader will need to be in a position to assess the justification for any claims that a particular conclusion can be applied to other situations. Finally, in such a practical field as ours, it will be important to weigh up the potential consequences of any outcomes reported for language learning and teaching practice.

In the subsequent **workbook** section the student is encouraged to see how this particular approach to appraisal might be used in practice on two complete research papers, before trying it out on their own independent reading. These personal readings of fictitious research papers comprise one worked sample appraisal, followed by a guided appraisal. Further guided appraisals are available on the author's webpage: www.ugr.es/~gporte.

The sample papers have been specially written for the book, but have objectives, method, data analysis, results, and conclusions that are based on a number of actual studies. The objective in writing these papers was to present them "warts and all": there has been no attempt to hide weaknesses or limitations in order to present "model" pieces of research for appraisal. Although the way a researcher expresses him- or herself on the page is a crucial element in communicating to the reader, adequacy of language or style of expression will not be specific objects of appraisal here, except where this reflects directly on the understanding or discussion of the research undertaken.

The appraisals begin with part of the text of a paper followed by leading questions, all of which are closely linked to the suggestions made in the relevant textbook section and which focus on an awareness-raising approach to critical reading.[4] In the worked sample appraisal, students can consider my own responses to these leading questions, alongside the ideas expressed in the

4. This initial pedagogical orientation towards critical reading owes much to the general advice and criteria offered by Bruce Tuckman in his book *Conducting Educational Research*, New York: Harcourt Brace College, 1994.

relevant textbook section. In the subsequent (guided) paper, readers should attempt their own responses based on the advice given in the textbook making use, if they wish, of the guidance provided by a number of additional specific prompts and suggestions.

As an initial recommended reading and response strategy for any paper read, and before embarking on these questions, the reader is encouraged to take notes on either side of the text, summarising, where appropriate, what has been read and then recording spontaneous reactions to it — much as if he or she were engaged in a dialogue with the researcher. An integral part of the approach to appraisal advocated here is that the reader begins to stimulate their critical faculties through such a stream of consciousness on the very first encounter with the text. Clearly, such spontaneous "feedback" may well need to be revised in the light of what is read elsewhere in the paper; however, both kinds of initial response will stand the reader in good stead as important references and points of departure for future discussion.

In the book, these spontaneous, initial reactions, and others noted in subsequent readings, are then assembled and formulated as observations (see below) interspersed throughout that text. These will provide suggestions which initially help alert the reader to ambiguous or incomplete information, potential flaws or inconsistencies, weak arguments, or simply points to follow up in the remaining text. The worked example aims to help the student appreciate the nature and degree of critical attention required during a reading and the kind of subsequent questions that might be posed as he or she is reading the rest of the paper. Clearly, both my responses to the leading questions and the observations made are merely suggestions for what could be passing through the mind of a critical reader; doubtless, readers themselves will be able to supply more of their own based on their experience and the textbook procedures.

As these initial activities imply, trying to approach a piece of published research using an awareness-raising strategy is very much about learning to become constantly responsive to what you are being told. This requires a different, more meticulous, approach to reading than you may be used to. I have attempted to convey this need for constant critical receptiveness during appraisal through these preliminary reading strategies and subsequent observations, thereby encouraging the reader regularly to attend to their "inner voice" during the reading. The objective is to help the student, in their own subsequent independent reading, to raise critical awareness sufficiently to enable them to pause, assimilate something they have just been told in the text, and form the pertinent response — either as a question or, as it were, an aside to

oneself for later use. This pause — as one judiciously approaches and responds to a section of the text — is visualised in the workbook samples as a numbered symbol (①, ②, etc.) in the text itself. In the worked appraisal, these symbols correspond to the kind of response and/or question I found myself asking as I read the present text. Subsequently, in the guided sample paper, the reader is encouraged to respond independently to the points raised.

In this sense, and in terms of actual class practice, the book lends itself to more than one method of exploitation. The approach should appeal to those who seek a way to combine teacher input, student input, and student interaction, and provide for different learning phases during and across classes. The textbook sections and worked sample material can be used as the basis for teacher input, group activity, and later discussion. The guided appraisal could be used as self-access material through which a student can work before submitting the responses to individual presentation, peer and/or teacher discussion.

The fact that the textbook chapters are ordered in the way they are is not to be seen as indicative of the way the book must be worked through. Rather, the units follow the format of a typical research paper in this genre. There is no intended progression as regards subject matter or complexity, although readers *are* advised to consider separately the section of the worked sample alongside the relevant textbook chapter before attempting their own appraisal of the guided sample and going on to their own independent reading. Doubtless, some readers will want to concentrate on the appraisal of certain sections of a paper more than others, and this is facilitated in the workbook samples by constant reference to what was addressed earlier in that particular paper.

Acknowledgements

Grateful acknowledgement is made to the following sources for permission to reprint material in this book.

Table adapted from Brown, J.D. 1992. Statistics as a foreign language: Part 2. *Tesol Quarterly*, 26, 4, pp. 629–664.

Flow chart reprinted from Hatch, E., and Lazaraton, A. 1991. *The Research Manual*, New York: Newbury House Publishers.

I also wish to record my thanks to the editors of the series for their encouragement and advice, to Kees Vaes of John Benjamins Publishing Company, and to Dr Phil Scholfield (University of Essex) and Professor Carl James (University of Wales), discerning reviewers who have also contributed much to this book with their observations and recommendations.

Graeme Porte

Textbook

1. Introduction

1.1 The abstract

For the author, the principal aim of the abstract is to summarise the most important points of the paper. This summary, however, will often be the first contact — and sometimes the only contact — that the reader will have with the paper. There are a number of different guidelines to authors about writing this, and other sections, of a research paper in our field. Typically, however, it will provide concise information and all-important indicators to the reader about what to expect in the body of the text.[1] From our point of view as potential consumers of the research, there are two main objectives: (1) We would like to have enough information provided to be in a position to judge if the study is sufficiently relevant to our own current interests to be subsequently read in its entirety *and* (2) where the study *is* going to be of interest, we may want to make a mental note at pertinent points during our reading of the abstract of aspects mentioned about which we would be looking for more details in the main body of the text. We might consider the importance to our research interests of the question asked: does it seem to have the potential to change, or add to, our theoretical understanding of the problem? This should also help us to make a preliminary evaluation of how successfully and appropriately the research has been carried out. Furthermore, we could be in a position to judge its initial merits based on our acquired knowledge of standard procedures in this kind of research. In order to help achieve these objectives, we could be looking to the abstract for some or all of the following:

---

1. Readers are reminded that the format and presentation of different sections of a research report may vary according to the requirements of the particular publication. Many journal editors in the field direct contributors to the guidelines available in the latest *Publication Manual of the American Psychological Association (APA)* (Washington, DC: Author, 1994) for specific information on the recommended detailed content of each section. Thus, where relevant, I will also refer to these recommendations in the body of the text.

a. a statement of the topic and aim of the paper, which may be accompanied by a statement more broadly situating the research.
b. a concise description of the sample[2] and materials used.
c. some information about the procedures used and the way data were later analysed.
d. a brief summary of results, or the general trend of these, and what conclusions are to be drawn from these.

1.2 The background to the problem and the problem statement

The introduction to the paper traditionally contains a number of sections whose aim is to establish a framework for the research in question so that we are aware of how it fits in with other research. While we should remember that different publishing media may suggest other content for, or sequences of, these sections, there will be a common objective to give the reader a clear idea about what is being done and why it has been done.

Is the background to the problem described? If so, what is it?

The background to the problem is the section which helps the reader to situate him- or herself in the area in which the problem is found. The section therefore might aim to rationalize the problem and explain why that problem is, in fact, a problem. This element will often be the initial standpoint of many research papers and can take its lead from, or refer directly to, prior published or unpublished work. However, this is not the "review of the literature" section of the paper — a subsequent, and comparatively more extensive, analysis of current thought on the problem. The writer might want to use this background section to set the situation for the forthcoming research in as succinct a way as possible and specify where exactly the problem has come from. This orientation is best accomplished by providing the background. One possible way to establish such a frame of reference for the research problem is to quote respected sources. Often, as a result of quoting these sources, the "conclusion" arrived at is that the problem has not been fully or sufficiently studied or that the present study, in some way, makes a useful contribution.

---

2. Terms in grey ink in the text are further explained in the "Glossary of key terms in quantitative research" (p.231)

Is there a problem statement? If so, what is it in your own words?

The next step in the introduction is often the problem statement. It is not obligatory to state the problem at such an early stage in the introduction, but one advantage of so doing is that the reader is thereby given a clear perspective from which to assess, firstly, the relevance of the paper to their work and, secondly, the subsequent arguments presented (in particular, during the review of the literature). Even if this has already been mentioned in the Abstract, it is useful to see specific identification of the problem here in the main body of the text. One or two sentences in the form of a clear statement might have been chosen to give us the idea. We might be on the look-out for sentences that begin: "Our main aim in this study…." or "The principal objective here…." or "In this paper I will describe/explore/investigate….", and so on. As an additional aid to comprehension of the problem, the problem statement might identify, where necessary, the nature of the principal **variables** studied, in particular **independent** and **dependent variables**, and perhaps suggest possible **interactions** between these. However, no great detail needs to be sought at this point beyond the concept itself, particularly since these variables will probably have yet to be operationalised formally within the study (see below, p.29).

One useful way of approaching the critique of this section of the paper is, first, to highlight the problem statement in the introduction and then follow its re-appearance throughout the text (including the Abstract itself). In other words, the reader should be able to find out the aim of the study and then confirm subsequent mentions as a check on this proposed aim. It should go without saying that, if there are various allusions to the problem statement at different points in the paper, these should all be consistent with, if not equal to, what has already been highlighted as the original problem statement. Any apparent discrepancies will need to be noted down, as they might well affect both the conclusions proposed and our appraisal of them.

Once the problem statement has been located and analysed in the above way, we will need to keep in mind — while we continue our reading — how far that statement, as expressed in the author's or our own words, is a comprehensive statement of what has been studied. For example, as we read through the rest of the paper from this point, we might feel that another slant begins to be taken on the problem, or that data are collected in a different way to that we had anticipated by reading the statement, or even that other variables are being presented for study, which do not appear in the problem statement. Yet the problem statement (and, particularly, the subsequent **research questions**)

should be taken as a comprehensive statement of aims. In other words, once an aim is proposed or problem statement made, the reader is right to assume that what he or she has just been told is going to be "the whole aim and nothing but the whole aim" — anything else could be seen as an afterthought by the researcher and, therefore, might affect our confidence in the comprehensiveness of the statement and, *ipso facto,* in the study itself. For example, it is particularly confusing for a reader interested in learning about the effect of a particular foreign language methodology on listening comprehension **scores** to discover that, although the problem statement promised such a comparison, the research question or **hypothesis** then goes on to analyse this **independent variable** against a number of *other* dependent variables — as well as listening comprehension scores. The problem with this is that the outcome might be a piece of research that provides a rather more broad, but less profound, study of a number of variables instead of the closer investigation of just the variable the reader was looking for in their reading of the paper.

These considerations about comprehensiveness will also lead us perforce eventually to reflect on the extent to which the problem was reflected correctly in the original statement. Again, this might sound like a "mere" question of linguistic accuracy, but it also reflects on the general confidence we can have in the study itself. For example, if the problem statement claims that "relationships" will be determined between two or more variables, it could both disappoint and frustrate an interested reader to find that what is actually being tested is the separate effect of the two variables named.

Thirdly, and as part of the appraisal, the expression of the problem statement should be seen to be unambiguous. An inability to comprehend what the proposed objective of a paper is at the start does not bode well for subsequent comprehension of the text. One recommended way of checking on the transparency of the statement is to try to put it into our own words and then ask ourselves whether — now in its rewritten form — we understand better the stated problem. As always during appraisal, any doubts created by such analysis need to be considered carefully. The fault is not necessarily on the part of the author, of course. Nevertheless, we will not be in a position adequately to appraise a piece of research if part, or all, of its stated aim is not completely clear to us at the start.

From the problem statement, do you understand: (a) the variables to be measured? and (b) the functions of these variables? If not, what values would you assign from what you have been told so far?

The issue of accuracy can also be extended to the correct assignment of the functions of these variables in the problem statement. To be in a position to understand how the variables in a study are supposed to relate to one another, we need to be clear about their proposed function — these functions are inevitably a central aspect of the problem itself and should, in turn, be clearly identified rather than being left to the reader to work out. A sound grasp of the functions of variables in quantitative research is particularly important to help us appraise how far appropriate statistical procedures have been carried out on data. For example, the widely-used descriptive or **inferential statistical** procedure known as the "*t*-**test**" requires a comparison of two **levels** (or groups) of only ONE independent variable. Thus, if a problem statement indicates that an independent variable "*native language*" is to be defined for the purposes of the study as "German" and "French", there are indeed two levels of the independent variable envisaged. However, if the same independent variable were defined as "subject-prominent", "topic prominent", or "mixed", we are being presented in the problem statement with a variable with *three* levels or groups. The **assumptions** of the "*t*-test", however, mean that only two levels or groups can be compared — no cross-comparing is possible. The reader might already have his or her attention drawn to potentially spurious findings before he or she even gets to the "Method and Procedures" section! Similarly, we might want to think ahead and consider whether we will expect the researcher to provide any **control variables**, based on what we have read (or know) about the research aims.

Since the variables have yet to be operationalised at this stage, we would be looking for their identification in a conceptual form only. Thus, the variables could be named, but no description of how they were measured is necessary at this point. However, as part of our awareness-raising approach to reading, we ourselves could begin to think of how the main variables might ideally be measured. Once again, this can help us to be in a better position to assess upcoming sections of the paper. For example, many descriptive or inferential statistical procedures require the variables to be measured in score (**interval** or **ordinal**) data form. If the problem statement appears to indicate (even though it does not mention overtly) that one or both variables are "frequencies", "percentages", "tallies", or "**ratios**" and that these data are to be used in their **raw** state (i.e., as **nominal data**, without conversion to another form), the reader

should already be aware of the possible inappropriateness of the statistical procedure about to be used.

Is there a contribution claimed to theory and to practice?

Finally, two further criteria are of great usefulness to our appraisal of the significance of the work about to be described. Judgements will need to be made about both the "Contribution to theory" and the "Contribution to practice" provided by the problem. "Contribution to theory" addresses the extent to which the outcomes have the potential i). to add to what we already know, ii). to help us better understand a particular observable fact, and/or iii). better evaluate a number of previous explanations or models. Therefore, it is important to remember that the assumed or stated contribution to theory is intended to provide an important link between research and theory and depends to a large extent on prior knowledge in the field. The author will need to make us more aware of this in the literature review, but at this point we might look for references to prior thought or previous theories that help us see where the current study actually fits in with what is already known. For example, if an author sets up an investigation into the effects of a particular ESL teaching context on L2 "acquisition" (i.e., rather than on "learning"),[3] he or she might be advised here initially to "place" this proposed aim against the past and current arguments about the proposed differences between what is thought to be "learning" and what is "acquisition" in second language learning. In this way, he or she will be able directly to suggest how far their own study's outcomes will throw further light on these arguments — and thereby contribute to current knowledge.

Ironically, such contribution to theory is to be welcomed, even when the stated aim of the paper is "only" to provide findings which are to be directly applied in the classroom. It is, perhaps, all too easy to forget — in today's pursuit of instant methods and quick solutions to classroom problems — that second language learning and teaching belongs to the field of applied linguistics which, in turn, means that research in that area inevitably aims to contribute to a particular science. It should be part of the professed aim of a descriptive, experimental, or quasi-experimental piece of research to contribute to the advancement of that science. Both theory and research are vital and necessary elements of science. Science without controlled, empirical research would

3. See Krashen, S. 1981. *Second language acquisition and second language learning*, Oxford: Pergamon Press.

consist of only untested ideas and biases. At the same time, science without theory would consist of a selection of disorganised observations and practices rather than meaningful comprehension of the psychological world. Advancement should mean — as far as possible — innovation and contribution at both a theoretical *and* practical level.

One way of assessing the "Contribution to practice", as the name suggests, is from the potential contribution provided by the study to promoting change or other kinds of application at classroom level. At this stage in the paper (i.e., before we read any of the assumed pedagogical implications from the point of view of the researcher), we might try to discover how the researcher predicts the outcomes of their study might have the potential to change the way the second language is learnt or taught. For example, the author may feel that their findings might provide greater insights into the success of a particular methodology and thereby help identify which elements of that methodology may be more successful in that context. Similarly, we might read about an hypothesised significant effect on language proficiency of a specially-designed language-laboratory programme. The author might then suggest that such results would support the implementation of the programme on a larger scale in classrooms.

As so often in the appraisal of papers, it will be important for the reader to be able to follow the researcher's line of theoretical and practical reasoning leading up to the problem, the research design, and the procedures used. In this way, we can also appreciate how far he or she anticipated certain results, and how far the contribution declared by the researcher in this section was later borne out by the findings obtained, which we will later be reading.

1.3 The review of the literature

Are you satisfied that the review describes the most relevant work done and indicates its relative importance?

Although there is no obligatory sequence of elements in the introductory section of a paper, we would now reasonably expect the author to fill in some of the background details previously sketched in through the "Background to the problem and the problem statement". The aim might be to use previous and current knowledge to continue to explore the problem, now from a variety of apposite angles and, by so doing, present the pertinent arguments which ultimately herald the study's hypotheses and/or research questions. This section

of a research paper — as opposed to a fully-developed thesis or dissertation — is often highly focussed and summarised. The main reason for this is that the author will have looked for key previous studies or relevant work that helps him or her eventually to underline the deficiencies or limitations in the current research area and the need for, and the importance of, the present study. This will have required a critical selection on the part of the author of work that, for example, highlights or discusses the same or similar variables to those about to be studied (according to the "Background to the problem and the problem statement") or suggests hypotheses that are to be tested in this study. The researcher takes on a particular responsibility in summarising relevant work for the reader and trying to make sense of the large amount of literature. Let us assume, for example, that we are reading a paper because we have been attracted to it by what was read in the abstract. Perhaps it promises new insights into a particular **construct**, a specific methodology, or a relationship of variables we are interested in. However, if every author of research were to start anew within that area, re-defining constructs or methodologies, working out entirely original meanings and definitions of variables, and then positing their own links between them, the resulting mass of knowledge would soon become confusing to the field — chaotic rather than summative knowledge.

Thus, one of the fundamental criteria we could use in our appraisal of this section is the perceived importance of the studies cited. Therefore, as readers, we will need to be attentive to references which, in some way, deal with the central variables or elements in the study and their relationship as declared in the problem statement. It would be logical for the reader to expect more detailed information about the studies which are more directly related to that currently being carried out and/or to the subsequent formulation of hypotheses. In other words, the reader should reasonably expect the review to be considerably more than a list of "what I have read"; papers are not just mentioned in the review for their own sake. Rather, we are looking for a summary of the literature, and past work should be cited for the light it potentially sheds on what is and is not known to the field and, thereby, on the author's *own* study. The advantage gained is that results of the most important work would then ideally be tied together so that their consequences for the present research become immediately clear to the reader. The author is then seen to be genuinely trying to carry on from a certain consensus reached by other researchers, and thereby profitably adding to the current body of knowledge. In this way, he or she builds a case or a context for the present research. Thoughtful organisation of the review of the literature will therefore enable the reader to follow logically, and thereby appraise, the questions emanating from

the background to the problem/problem statement and the arguments leading up to the subsequent research questions and/or hypotheses themselves.

It is also useful for the reader to have clarification on the relevance and relative importance of the work cited. Unfortunately, there also exist literature reviews which are little more than disguised bibliographical lists. This is often signalled by padding, or what appears to be the citing of particular researchers or work just for the sake of it. Such a procedure may be a warning to the reader that little relevant or reflective background reading has gone into building up the contribution of previous research to the study. Within this criterion of appraisal, the reader should also be made aware of the comparative importance of any results from the studies reviewed. Inevitably, some results will have more bearing on the current problem than others and the reader needs to be aware of this as he or she digests the previous knowledge available on the subject, thereby to be in a better position to judge where the present study fits in. It is also useful for the reader to try to determine, as far as possible, whether any relevant work has been omitted.

Are you satisfied that the review has sufficient critical address of the literature?

Such explanation on the part of the author is not limited merely to summarising what is already known, of course. The critical consumer of this research could reasonably expect the researcher, in turn, to have been suitably discerning and analytical in their reading of this selected literature. It is instructive to read a paper which not only identifies, but also explains and critically addresses issues relevant to a study that have emerged from this reading of the literature. We might usefully see such critical engagement on the part of the author presented in the form of a structured explanation showing what the author has found satisfactory or wanting in an argument or finding — rather than mere unsubstantiated declaration. Progress in our science will very much depend on how researchers review the work of their predecessors and make suggestions about how to improve or change things. Through such description and explanation, the author might be able to produce a concept or build a theoretical structure that may explain facts and the relationships between them. The importance of theory, as we have seen in the previous section, is both to help the researcher summarise previous information *and* guide their future course of action — thereby establishing the contribution to theory provided by the study. Often the formulation of such theory through a critical review of the literature will help indicate missing ideas or links and the kind of additional data to be provided by the study in question.

The nature of this critical address will be another factor in assessing the amount of confidence we may feel able to place in the paper. Much will depend, therefore, on the way the author argues his or her case before us. We will return in Chapter 4 to the way the reader needs to analyse findings put forward in discussion and conclusion sections of published research. Suffice to say here that the reader could hope the author indulged in a similar practice in his or her own reading. Such critical address will probably not need to be applied to each and every study cited in the review, of course. Much of the previous research related to the background of the study, for instance, may require little more than description of the main outcomes. However, we might want to know more about how the author interprets findings or conclusions which have a direct bearing on the problem statement itself. Ethical standards in reporting of research mean that such address should not be dismissive or aggressive towards previously published work: good scholarship requires author (and reader!) to adopt a positive and enquiring attitude to ideas presented by others. Such an attitude has the advantage of transmitting to the reader the idea that the author is able not only to analyse previous arguments and findings in a way that synthesises this work, but also to use this synthesis to generate new ideas and open up research directions. Typically, we might appreciate a stance which shows us both what the author has found acceptable or unacceptable in a particular idea or finding *and* which then also explains why the author has reacted in this way.

It is often useful for the reader, to whom the results of the author's literature search are now being presented, to get into the habit of appraising how authors classify their findings, and how they critically explore and explain relationships between findings. Of particular interest will be the treatment given to potentially conflicting findings and controversies within and across studies: such conflict is quite common in research in our area and, where it is found, we might reasonably expect the author to explain, or seek to provide possible insights into, such deficiencies in our knowledge or discrepancies across studies. There is always a temptation to ignore such differences in the interests of summary; however, these should be of interest to the researcher *and* the reader, as such apparently inconsistent findings can actually provide a rationale for further study. For example, the researcher may subsequently decide to examine the same question from a different point of view, possibly with improved methodology, so as to offer more convincing findings. For the reader, learning about areas of knowledge where findings are unconvincing or in some way incomplete can also help him or her to form their own research agendas based

on current needs in the field. Furthermore, the common practice of "evening out" findings in the literature review will certainly mean that we lose information, and it may mean the possible complexities of a problem are not fully grasped — or communicated to the reader. The literature review should transmit the idea that the author has studied existing work in the field with analytical insight, leaving us with a balanced picture, albeit necessarily partial, of the current state of knowledge and of major questions yet to be answered in the relevant subject area.

Are you satisfied that the review communicates the main points related both to the background to the problem and the problem statement/ independent and dependent variables?

Once again, there are no explicit rules about the way a review should communicate the main points to the reader, but our reading and appraisal of the review can be facilitated if there is logical progress within the text itself. The reader's understanding of the situation that led up to the research can be enhanced if there is an evident movement from reviewing general work within the background to the problem towards more specific research which reflects on the eventual statement of the problem itself and any research questions or hypotheses to follow. Within this coherent progression, it would be interesting to read about the kind of methodological approaches or analysis of data used by previous researchers in the area, together with how and why the present author seeks to replicate, or improve on, these. Another useful way of helping the reader appraise how the literature review reflects on the study about to be described is to move from studying the relevant literature on the main independent variable(s), to that done on the dependent variable(s), and finally concentrating on the literature which has related the independent and dependent variable(s). Other problem statements might suggest a chronological approach to reporting the literature.[4] Whatever the way chosen by the author, the idea behind any of these "routes" is gradually to focus on the situation in question and pave the way for the research hypotheses or questions. As readers, however, we will want to be able to follow the route chosen throughout the text, and this

4. Other approaches to writing this section have been suggested. Slavin (1986), for example, recommends a "best-evidence" approach to the writing of this particular section, whereby strict criteria are established before a study is included for review. (Slavin, R. "Best evidence synthesis: an alternative to meta-analytic and traditional reviews". *Educational Researcher*, 1986, 15, (9) 5–11).

means continually stopping and asking ourselves if we understand the arguments put forward so far. One way this can be done is by attempting to recap the main points made in each paragraph. If we can do this in our own words, we might assume that the author has successfully communicated what needed to be said. Our own summary obtained should likewise reflect logical progress from the general to the specific.

Are you satisfied that the review covers an adequate time-span?

The reader will also want to consider the extent to which the literature read and/or studied adequately reflects current thinking on the subject and covers an adequate time-scale. There is little to be gained by an author including research that has now been superseded and/or contributes little or nothing to the current argument: this may result in mere "padding". The problem here is that the reader may not be adequately informed of the research carried out. Nevertheless, we are right to want to be made aware of the latest, and most relevant, information which reflects on what we are about to be told. This does not necessarily mean that the review should include only the most recent work in the area. Indeed, references may be found to many "classic" studies in the field that are over twenty-five years old. The key here is that we are looking for previous evidence that helps us see how the current problem or situation has come about and thereby clarifies the need for the current work about to be described. Indeed, even if the author claims to be dealing with a relatively virgin area of knowledge, it will be advantageous for us to see how the apparent deficiency in the field has come to light as a result of past and present research. The critical reader might expect at least an explanation of the "route" by which this new ground was reached. Inevitably, this will mean the author giving information about how and where their study fits into the field and the current body of knowledge and also, *ipso facto*, how this particular study came about.

Are you satisfied that the review has adequate reference, where necessary, to empirical work?

In much of the experimental and quasi-experimental research in our area, we will also want to be concerned about how far the work being critically addressed here has not only been directed to similar concerns as the paper in question, but also has used similar empirical methods. This comes down to our assessing how far a fair comparison is being made between like and like. For example, if the author of a paper wishes to convince us of the need to provide more experimental data about the order in which second-language learners acquire their L2

vocabulary, we might logically expect that some or all of the literature addressed includes findings which are based on data gathered with similar aims rather than merely derived from discursive literature about the way these students might acquire their vocabulary.

Does the review succeed in convincing you of the need for the study?

As a result of our reading, we should be able to comment on how convincing the review is: in other words, has the author succeeded in making a strong enough case for the study in question? In particular, the reader will need to ask him- or herself whether meaningful, significant, and innovative data are likely to be contributed to the body of knowledge as described in the review.

1.4 Research questions and hypotheses, variables, and operational definitions

Our route through a sound research paper is typically well-signposted by the researcher to ensure that we are accompanying him or her on the same road and that we all understand the direction in which we are going. So far there have been a number of key "signposts" in the text which have guided us. The abstract should have given a clear idea of the route to be taken, then the problem statement confirmed the main direction of the study. Now — in the light of what has been revealed from the review of the literature — the final "signpost" can help to refine the problem statement and clarify the objective as research questions or hypotheses.

Are research questions or hypotheses formulated? If so, what are they?

Research questions and research hypotheses are different, and the reader should also be able to appreciate that difference in the method and procedures to be followed in the rest of the study. **The research question** is more typical of descriptive and survey-type research; it does not need to predict any possible outcome as such. The main idea is to look into a particular situation and "see what is there". The result may well contribute to generating hypotheses for future testing, but this is not obligatory.

Usually when we have questions, we want to know the answers. But this does not mean that we have no idea of what those answers might actually be. After thinking about the question ourselves, reviewing the literature, and studying what others have discovered about the questions we ask, we may come

up with suggestions about possible answers. These tentative answers, written formally, are research hypotheses. **The research hypothesis**, most often found in experimental and quasi-experimental studies, does provide a suggested response or expected outcome to the problem described in the problem statement, previously outlined and discussed in the review of the literature. Basically, this is a hunch that the researcher has about the existence of relationships or differences between the variables used and which will be subject to examination through the subsequent investigation. Indeed, the way this relationship or difference is perceived in the hypothesis provides a guide to the researcher (and the reader) as to how the original hunch is to be tested among the "subjects" in the study. Thus, the reader will need to attend closely to the research hypothesis, since what follows will need to be structured in such a way as to enable the stated hypothesis to be tested.

Typically, the research questions or hypothesis would be located after the literature review, wherein logical and empirical support for such statements would hopefully have been found. Once read, it is always a good idea to underline the hypothesis or questions to be able to refer back when reading and interpreting the results section. Since these hypotheses or questions give the study direction and form, our understanding can be enhanced by explicit statements of such questions and/or hypotheses and not mere implicit objectives, which leave the research question or hypothesis and subsequent results open to whatever interpretation. Particularly in the case of **directional hypotheses** (see below), look for explicitness and confidence through the use of words such as "*Results were expected to confirm that……*", or "*Data collected should support the notion that……*", and so on. It should be clear from the above that the research questions and/or hypotheses (in whatever form they are introduced) represent a key element for a reader trying to understand the connection between the perceived problem and the way the author has chosen to go about responding to it.

Are the research questions exploratory, descriptive, or explanatory?

Very often the researcher's prior study of the field and review of the literature will have exposed a need to explore, describe, or explain further a particular phenomenon through research questions, *before* arriving at possible hypotheses. It will be useful for the reader to predict the nature of the study suggested by the particular research question as part of the valuable practice of progressively building up a critical response to what he or she is being told. In this way, we can begin to envisage outcomes and perhaps already consider possible problems

or drawbacks in the research design. Exploration will see a research question in which the researcher aims to find out what is happening, to seek new insight, to pose new questions, or to attempt to assess the phenomenon in a new light. In other words, the study structures the research rather than the other way round and the research may thereby become one of hypothesis building rather than **hypothesis testing**. Descriptive research questions will attempt to portray an accurate profile of people, events, or situations. Finally, explanatory questions will seek an explanation of a situation or problem, often in the form of **causal relationships**. A particular study may be concerned with a combination of all three tendencies, but we should be able to highlight one principal trend through our initial reading of the research questions(s).

Are the hypotheses offered directional and do they predict differences or relationships between variables?

As one might expect when a hunch is at the centre of the argument, the hypothesis itself is often couched in language of prediction, which sounds as if the author is making a bet with the reader (e.g., "*The number of errors made by elementary German-as-a-foreign-language students is related to the kind of methodology they receive*" or "*The ability to discriminate between minimal pairs in L2 Portuguese increases with age and educational level*"). The hypothesis should ideally present the following information to the reader: firstly, there should be some statement concerning assumed relationships (or lack of them) between the variables, or the presumed influence of one (or more) of the variables on the other. Secondly, we can expect to read an hypothesis which really can be tested: that is, it looks to us as though it will become possible to assign **operational definitions** to the constructs or variables described and thereby produce useful data which can then be analysed.

When a researcher presents an hypothesis, this will traditionally be written in either a positive, directional, or a null form. The choice of presentation is important for the critical reader too, for it has repercussions both on the way we will expect subsequent data to be statistically analysed and on the way findings should be interpreted. Any hypothesis will represent a written form of the "hunch" that the researcher has about the outcome of the study. A **positive** hypothesis might declare that there will be differences found as a result of the interaction between the variables but does not specify whether these will be positive or negative outcomes. A **directional** hypothesis goes one step further by informing the reader about the specific trend of the difference or relationship. For example, we might read that instruction in certain reading comprehension

strategies is hypothesised to result in *fewer* errors of comprehension and/or *better* acquisition of vocabulary. These hypotheses go beyond merely saying "something will happen" and state exactly what will happen (i.e., there will be *less* of something and *more* of something else). However, such direction cannot be an arbitrary decision. Direction in an hypothesis should be understood by the reader to proceed logically from the data provided in the review of the literature. In other words, other researchers may have already revealed or suggested a particular direction using these variables in other contexts. On the basis of previous research in the field or previous theoretical discussion, the researcher is thereby prepared to predict that a relationship will exist between the current variables *and* that such prior knowledge also permits him or her to predict the direction of such a relationship.

We will need to reflect back on what we have read in the literature review as we take in such an hypothesis. However intuitively acceptable a directional hypothesis may sound, if it is not soundly based on previous outcomes or theory, such an hypothesis may become impossibly vague. In our example of comprehension strategies above, for instance, fewer comprehension errors may indeed be an effect of the teaching; nonetheless, a number of other explanations for such outcomes may also be possible. It may be something to do with the speed with which the subjects read, or the difficulty of the passage used, or any number of variables. When previous research gives us a clue to what might happen, however, the researcher has considerably more empirical evidence to back up any such claim.

In the field of second language learning, with its many differing teaching and learning contexts, methodologies, objectives, and students themselves, it is always going to be difficult to support directional hypotheses beyond reasonable doubt. Much previous literature may indeed exist, and other prior studies may have experimented with, and confirmed, similar outcomes using these variables. Conversely, it is more often the case that relatively little replication is done in our field and the researcher may often have few valid equivalent studies to support their own findings. Given such a situation, it is hardly surprising that many experimental and quasi-experimental studies in our field opt for presentation of the hypothesis in its **null** form: in our example, the null version would read "There is no effect of instruction in certain reading comprehension strategies on improvement in comprehension or vocabulary acquisition". If this **null hypothesis** is supported, then instruction will have no effect on comprehension or vocabulary acquisition — hardly what the serious L2 teacher would be hoping for! If the null hypothesis proves to be incorrect, then such instruction

might prove useful. In this case, the researcher may be using descriptive or inferential statistical procedures (see Chapters 2 and 3) which help him or her to reject the null hypothesis in favour of the logical alternative. Use of the null hypothesis has become traditional in experimental and quasi-experimental research in our field because it allows the researcher working in what are such heterogeneous language-learning contexts to be conservative in their initial "hunch" and more cautious in the conclusions he or she draws from the data. Indeed, analysis of the caution expressed in extracting conclusions from results based on such hypotheses will be one of the elements in our appraisal of the discussion and conclusions of such a paper.

The problem is that hypotheses can be dangerous things if used carelessly. As cautious readers of research, we will need to be attentive to what is read into any possible outcomes. First and foremost, and in order to carry out any test of an experimental or quasi-experimental hypothesis, it must in principle be possible for the predicted effects either to occur or *not* occur. Therefore, to return to our example above, it must be possible for the reading comprehension strategy instruction either to bring about improvement in comprehension or vocabulary acquisition or not to. This is a basic tenet of experimental research. If there is no chance that a particular experiment will not go the way the researcher predicts, then there may be little point in doing the experiment. This is another reason why the null hypothesis has acquired its traditional use: this states that the researcher will *not* find the results he or she expects.

Secondly, we may often come across a piece of research which attempts to use the results from a directional hypothesis or a rejected null hypothesis to suggest that a particular effect of, or relationship between, variables is thereby proven. As readers of research in L2 learning, it is worth recalling that most study in our field is perforce exploratory in nature. It is unlikely — and to a certain extent, undesirable — that a study starts out to *prove* that an hypothesis is or is not correct. Although certain research designs and inferential statistical procedures will give a researcher the possibility to generalise from their findings to other contexts (see Chapter 2), the process should be interpreted by the researcher as one in which support (rather than outright proof) is sought and found for their hypotheses. The reader will need to remember this as he or she later reflects on the findings and discussion presented. Confirmation or rejection of the original hypothesis is unlikely to lead to any firm conclusion or provide definitive answers to the original problem statement. Much more desirable is that what comes out of the research is destined to become another element in the current body of knowledge which, in turn, will be used to refine

a prior theory, provide additional evidence for a phenomenon, or open up further research directions.

A further useful exercise for the reader at this point would be independently to consider any alternative relationships between variables not stated but implied in the previous review of the literature and the problem statement. The researcher may have decided to limit the study to only one aspect of the problem and presented only this in the research questions or hypothesis. Nevertheless, the reader might also have appreciated during their reading other possible relationships between the variables that could be used to shed light on certain findings in the study. It will always be useful to note down somewhere any such perceived alternative relationships or explanations in order to approach the results and discussion section with a more critical eye and also to consider how far the author has successfully, convincingly, and completely explained the findings obtained.

Are the research questions/hypotheses unambiguous, consistent with the problem statement, feasible, and supported by the review of the literature?

Once we have located and established a research hypothesis or question, the reader might usefully make a judicious pause in their reading in order to form an initial response to this defined aim of the study with the use of a specific set of criteria. This break will ideally allow us to ponder the potential consequences — as we see them — for the forthcoming method and procedures of this expressed aim. It will further help guide our reading of upcoming sections of the paper and, once again, prove an important element in establishing our confidence in the study itself.

In the previous section I mentioned how, ideally, the literature review should guide the reader through the background to the problem and "dovetail" into the research questions or hypothesis by way of the critical engagement with prior studies or with current knowledge (or lack of it) in the field. Thus, the reader should then be able to use the research question or hypothesis, as the researcher does, to focus the study, to give it direction, and to make it easier to follow. As such, the hypothesis or research questions should ideally be unambiguously stated in the body of the text. Although hypotheses are abstract and ultimately concerned with theories and concepts, this does not mean that the research hypothesis itself should be reported in abstract *language*. Once again, it is useful to have the hypothesis (or research question) re-stated in the reader's own words. Such rephrasing can often help us to make note of key concepts or constructs that suggest the need for adequate definition and/or further explanation,

perhaps as a consequence of the imprecise language used by the author. It should be clear to us what is being hypothesised to relate to what or which variable is thought to affect another.

Consistency can be ascertained by checking whether the hypothesis (or research question) presents us with the kind of likely outcomes that will provide adequate responses to what has previously been posited in the problem statement and subsequently analysed in the review of the literature. The review of the literature will ideally have already provided the basis for previous support for the hypothesis or research question, thereby convincing the reader that this is based on substantially more than mere intuition on the part of the author (or, worse, one made up after the researcher had seen the results!).

It would also be wise for us to start thinking about a number of factors with respect to the perceived feasibility of the study as we understand it so far. If no time interval is mentioned as yet, we might want to consider how far the researcher could reasonably accomplish the aims in the question or hypothesis as regards the time and access to subjects apparently needed. For example, if a researcher's expressed aim is to describe the language development of a group of "bilingual" children (because their mothers talk to them in one language and their fathers in another), we might suggest it would take years (if not a lifetime) adequately to complete the study. If the researcher defines the time period involved, we might think about the advantages and disadvantages of this and any constraints on possible outcomes. We then need to envisage the extent to which the researcher might reasonably have sufficient contact or access to a large enough sample of subjects or, indeed, whether that access will be of sufficient value to permit useful data to be obtained. We might also look carefully at the number of variables involved or the complexity of the relationships expressed to consider how initially workable it all looks as a research design. In short, thinking about feasibility will help the reader to reflect on how far he or she regards the research question or hypothesis as manageable given the circumstances described so far and, therefore, what kind of constructive outcomes may be possible. It may well be that further reading of the upcoming "Method and Procedures" sections will help clear up any doubts we have on these points; nevertheless, forming our own opinions now will help us to be in a better position to appraise results and discussions later on.

A research question or an hypothesis is likely to have been justified or supported in one, or both, of two possible ways. Firstly, there is the route via logic. A theory will allow for a large number of potential associations between its component parts. The author may have described and/or analysed such a

theory and its constituent elements (see above "The Review of the Literature"). Any number of possible relationships to be tested will have been revealed and from which outcomes will — in sum — provide a testing ground for validating (or rejecting) that theory. On the other hand, empirical justification requires reference to previous studies and findings. We would have expected the researcher to have conducted a review of the relevant literature, appropriately reporting on, and summarising, the significant prior published studies that reflect on the supposed description or link between variables and that support this logically and/or empirically.

Whichever route was chosen, the researcher would want to have provided the reader with enough justification or support for each hypothesis or research question to assure us of its reasonableness and soundness. Published research may be encountered wherein the author has assumed too much on the part of the reader or taken for granted the fact that their arguments have led logically to the research questions or hypothesis. If there is no underlying reason for, or background to, a research question or hypothesis, the reader is being left to work out from whence that question or hypothesis came. A normal reaction might then be, at best, perplexity (the reader is not able to appreciate what led up to such a statement in the first place) or, at worst, scepticism (the reader cannot appreciate how variables can ever have been thought to be related in this way). The wary or confused reader may even decide to dismiss the study's findings completely because he or she feels that the author has simply "fitted" the data into a post-conceived hypothesis or research question, according to the results found.

> **Can you identify the principal variables of the study, and are these to be measured as nominal, ordinal, or interval scales? Comment on the perceived appropriateness of these scales. Are moderator or control variables evident?**

It is important that the reader be able clearly to identify which variables are which in the study, to understand the roles these variables will play in the research, and, ideally, how these variables are to be operationally observed. This is useful because it will help us later to determine how far the data have been correctly analysed and interpreted — and therefore how much confidence we can place in the conclusions themselves. Obviously, if the research hypothesis or question is not stated clearly enough in this respect, the reader will have little basis for subsequent judgement. Independent and dependent variables should be obvious although, in certain studies (see below), such classification is of lesser importance.

In the kind of empirical research we are concerned with here, it is helpful to be able to appreciate what the independent and dependent variable is, together with any moderating, control, or possible **intervening** variables. Summarising, the *independent* variable is the element that the researcher believes may in some way relate to, or influence, the dependent variable. The *dependent* variable is the major variable that will be measured or observed to determine how, and if, it is affected by the presence of the independent variable. We expect performance on this variable to be influenced by the other variables in the study. In other words, what the researcher may be suggesting in his or her hypothesis or research question is "My hunch is that data from my dependent variable will be related to or affected by the way my independent variable is used". While here I refer to single dependent and independent variables, it is important to remember that studies can deal with many variables, which may be further independent/dependent or **moderator**/control variables. The key, however, to understanding what the researcher is attempting to do is to perceive the "link" between the variables. A research hypothesis may suggest that a change brought about on the dependent variable will have its origin in the independent variable or the way that variable is used or manipulated during the study. It follows that a researcher may manipulate the independent variable as well as measure it, but the dependent variable itself will only be measured — it should not be manipulated.

In other studies, we will understand from the hypothesis that the researcher is less interested in predicting differences as a result of manipulating the independent variable, but rather in investigating the extent to which the variables are related. For example, the researcher might want to know whether one group of students score similarly on one kind of listening comprehension test, (e.g., listening to a tape recording in class) as on another (e.g., listening to, and watching, a video recording). On this occasion, neither of the variables need be denominated independent or dependent; both will be measured (rather than one of them manipulated) in order to see whether students who score highly on one test perform similarly on the other test. For the reader, the difference is another important one to establish, as statistical tests vary for testing differences and testing relationships. Furthermore, since neither of the variables is being manipulated, it will be impossible for the researcher to predict or suggest which variable is having an effect on another. All he or she will be able to state is that a relationship does or does not exist; any number of explanations for that relationship could be ventured (and, later, tested), but are not consequences of this particular experiment.

It is also worthwhile, when reading the research question and/or hypothesis, to distinguish between variables themselves and levels of these variables. For example, we might want to know how well L2 students are able to do a listening comprehension test, comparing when they hear it read on a tape and when they hear it read by a native speaker in class. The variable here is "L2 student", but that variable could then be divided into *levels* for the purposes of the study. If the subjects were of different nationalities, for example, a subject might be subclassified as "German", "French", or "Spanish", so that posterior comparisons could be made between *levels of L2 student.* The original variable now has three levels. Such a distinction between variables and levels of a variable is rarely made explicit, and so the reader may have to work this out. It is, however, important to do so at this point in the reading, since the function and measurement of variables will determine, for example, exactly what kinds of statistical test will be appropriate when analysing the data.

The *moderator* variable describes a particular type of independent variable that is thought to mediate or moderate the link between the "main" independent and dependent variables. Thus, the question being asked here is something like "I want to see what will happen to my dependent variable when it is affected by my independent variable, but I suspect that that relationship is also affected in some way by the presence of another factor — the moderator variable". As we might imagine in our field of L2 learning, the cause and effect relationships between an independent and dependent variable are relatively complex. A number of factors might mediate between the two to produce the outcomes described. Often the link between the two remains ambiguous until such a moderator variable is identified. Such identification will often be made by the researcher after reading findings from previous studies, as a result of which he or she wishes to study the effect of such a variable on the main independent-dependent relationship.

Similarly, the researcher — perhaps as a result of reading previous studies — may be interested in reducing or counteracting the effect of certain elements in a study to make sure they do not have a moderating effect on the relationship to be studied. *Control* variables are factors which the researcher deliberately decides to control in order to cancel out any possible effects on the main relationship studied. Thus, the researcher will not actually be studying such variables (unlike the study which *is* made of the moderator variables); the effects of control variables are merely offset.

In such circumstances, however, it will be important to consider the care perceived to have been taken in such control. We will very often only discover

which variables have been deliberately controlled for in the "Method and Procedures" section of the paper. So, and not for the first time, it is wise to think ahead and make a mental note of any reasons which we may have thought of during our reading of the paper so far for controlling one or other factor. We should then be in a better position — come the subsequent section — to decide on the appropriateness of the researcher's own judgements. Such decisions on our part and on that of the researcher allow us to perceive patterns in the data. Learn to become aware of the importance of variables such as age, sex, intelligence, previous knowledge, language ability, or dropout rate, each of which may or may not have been controlled for — with the resulting consequences for any findings revealed.

Without controls, patterns in data are often less easy to perceive. Nevertheless, we will need to recall that — however expedient controlling certain elements may be for the research design itself — such a procedure imposes its own "control" on the interpretation of findings. Depending on the kind of control exercised, the researcher will not be able to generalise very far (if at all) in the conclusions or discussion beyond the current context and the controls imposed therein. Indeed, the use of controls in studies lends itself to further study — whatever the results — as future researchers gradually release any previous controls in order to perceive the outcome. This is something to look for as we read the "Further Research" section often found towards the end of a paper: the present researcher (or we ourselves) might suggest which controls could be dropped or eased in the future with a view to improving the generalisability of findings. This contributes to further discovery by allowing researchers to find out which controls most influence outcomes.

Identification and function of variables are not the only pieces of information the reader will need to look for in this section of the text. Equally important is to be informed of the way the main variables are to be measured and to begin to think about the appropriateness of such measurements. Although more detailed information will invariably be presented in the "Method and Procedures" section of most papers, initial indications of measurement are often found here in the operational definitions of variables (see below). If no information is offered at this point, it is still a worthwhile exercise for the reader to envisage how each variable might reasonably be measured. This helps us raise our critical awareness enough to be in a better position to evaluate the appropriateness of the kind of analysis suggested here by the hypothesis or research question and later described in subsequent sections of the paper. A *nominal* scale gives a name or category to something. Measurement involves classifying

the data according to the presence or absence of that attribute — it might be sex, nationality, proficiency level, and so on. When we begin to tally how many objects or people or subjects have this attribute, we end up with a **frequency** measurement. It is important for the critical reader to appreciate the limitations of such data for posterior analysis. They can be very useful if we want to know, say, how many men and how many women took an exam or how many of these were intermediate level, advanced level, and so on (assuming, of course, that "intermediate" and "advanced" were operationally defined). On the other hand, if we want to do any real comparison between group performances involving mathematical manipulation, unconverted **nominal data** are of limited use.

An *ordinal* scale puts the data in order or ranks. Very often we will come across such data measurement in papers where opinions or attitudes are collected in terms of increasing or decreasing numbers or definitions on a scale: 1 through to 5 or "totally in agreement", "partially in agreement", "neither in agreement nor disagreement" through to "totally in disagreement". For the reader, it is again important to assess the value of such data for analysis. Although we do end up with important information about **rank order**, and this order has useful arithmetic worth, the value assigned is not precise. Thus, it does not follow, for example, that there is an equal mathematical unit or interval between the judgments of "totally in agreement" and "partially in agreement" or between this and "neither in agreement nor disagreement". This problem can be got round by increasing the number of alternatives from which to choose from five to, say, seven. The researcher then treats such interval differences as equal because the increased number of options is said to encourage the respondent to think in equal-interval terms. As I explain below, readers will have to make up their minds on the appropriateness of this kind of procedure.

Finally, *interval* data not only order or rank, but do so by also reflecting that rank as points on a scale. Furthermore, each interval is assumed to have the same value so that units can be added or subtracted and also subjected to other kinds of mathematical analysis. Indeed, very often, interval data like these are drawn from some kind of test or exam score, but we will still have to think about the extent to which the test itself encourages equal-interval data. For example, in equal-interval data, the difference between obtaining scores of between 1% and 3% in a reading comprehension test might be thought to be the same as that between 98% and 100%. Both differences amount to two "percentage points", of course, but we might wonder whether students scoring such high scores on a test do so because they are obtaining their "points" by

getting rather more difficult answers right than those who score between 1% and 3%. If, after reading about the kind of test involved, it is felt that the interval data collected may be seriously affected by the reliability or validity of the instrument of testing itself (see Chapter 2), it might be suggested that the researcher would have been more justified in using ordinal data analysis.

These mathematical and logical constraints often placed on data measurement mean that we may also come across research wherein data are overtly or covertly converted from one type of measurement to another, presumably to allow for better, or more insightful, analysis to be carried out. If such conversion is suggested either at this point or in later sections of the paper, the reader might want to consider (or be told) how appropriate such a conversion is and the reasons for it. The idea is to think about what information might be lost in converting interval to ordinal data or what information may actually be distorted by converting, say, ranked scores on a listening test into four (nominally-designated) groups of "excellent", "good", "fair", and "poor" performers. The argument is that we still end up looking at the same data, but from different directions. However, in the latter case, a consequence of the conversion is that we find ourselves comparing four grouped performances rather than individual rankings between students. This will inevitably affect the kind of conclusions to be drawn from the data, and the reader will need to be aware of this as he or she reads those conclusions.

Can you predict any intervening variables or contributory factors — if not stated here — that might affect findings?

Independent, moderator, and control variables can all be manipulated by the researcher and the effects on the dependent variable of such manipulation can be observed and/or measured. Thus, these variables can all be identified prior to the study itself and form an integral part of the research design. However, sometimes the researcher will obtain data from the original hypothesis wherein the effects of the independent and/or moderator variable are not clear. We often hope to establish some sort of direct link between independent and dependent variables in a study. In planning their research, ideally the researcher would have been able to identify all the important variables (or control for them). However, sometimes this is impossible. An *intervening* variable is a factor that — in the light of findings — has now been judged to have affected the original hypothesised relationship. How, and to what extent, it affects the original link between the variables may only be revealed by future detailed study of the effects of the "main" independent and moderator variables on the dependent variable.

In essence, the moderator and intervening variable look at similar things; the difference is that the latter is not directly observable and therefore has not been or cannot be identified precisely for inclusion or control in the original research.

This is not a fault on the part of the researcher of course, but it does have important consequences for the reader of the paper. It means that we must be aware during our reading that the author may not have been able to identify all the important variables for us and, perhaps, control for them. Indeed, in our research field, this may frequently be the case, for we are often studying internal mental processes that we may or may not be able to measure accurately. Factors such as intelligence or test-taking ability may not be directly observable or measurable but may affect research outcomes considerably. For the reader appraising the paper, however, it is important to begin to think about what may contribute to an observed effect/relationship or lack of it *before* being presented with the actual findings. Such a prediction can often be made by reading through the hypothesis and thinking "Is there anything which forms part of that (independent) variable as it is described here which might affect the hypothesised results?". In other words, look closely at the (independent) variable and try to work out what process (inside the subjects, as it were) might be going on to explain the notional outcomes.

Beginning to think about possible intervening variables or indeed other, more detectable, factors that might affect outcomes is far from being a pointless exercise for the critical reader. As so often throughout our appraisal, we are trying to make an active contribution to the text we are reading, as we are reading it. Identifying potential problems such as these can set us off thinking about why a certain outcome may occur and help us identify further research areas or studies which may provide more answers by using this information as their starting point. Moreover, if the researcher has not been able to identify all the important variables or contributory factors in a supposed relationship, we will need to remember in our evaluation of the results that there may be inherent sources of "error" in these data. For example, in a cause-effect relationship, this may be because only a certain number of the outcomes described in the dependent variable were due to the **main effect** of the independent variable.

If there is no explicit identification of variables in the problem statement, research question, or hypothesis — and particularly if it is the kind of study where these *should* be identified — the reader might look forward to the method or analysis section of the paper to glean further information. Thus, for example, particular statistical analyses presuppose certain denominations of variables. The commonly used *t*-test, for example, investigates the effect of two

levels of one independent variable on one dependent variable. In the procedure known as *One-way ANOVA* there should have been only one dependent variable and only one independent variable with three or more levels. Names of variables can often be discovered by looking at results tables or from tables of means and standard deviations in the "Results" section.

Although the identification of variables by the reader is an analytic procedure, rather than an evaluative one, it is a process which helps us understand what a study is about. On certain occasions, however, it can help to reveal additional insights into the data offered later and/or to weigh up any findings. One of the most serious cases is when confounded research designs are suspected. Here, the design of the study has made it impossible for the real effects of the independent variable to be measured across groups, since elements of one variable are present in more than one of the groups examined (see Chapter 2).

> **What were the constructs used and have these been adequately delineated to permit operational definition? How have these constructs then been defined operationally, where necessary, and is this description acceptable in its present form?**

Researching in the area of second language learning inevitably means working in such well-established and defined areas as *bilingualism*, *motivation*, *language proficiency*, and so on. Such areas or constructs are, however, very broad, and previous researchers would normally have facilitated the present researcher's work by narrowing them down in certain ways and making them much more precise or specific. Thus, for example, *motivation* is a construct which has been focussed down to *intrinsic*, *extrinsic*, *instrumental*, *integrative*, and so on. *Language proficiency* has been further defined as *Advanced*, *Upper-intermediate, Intermediate, Post-beginners,* and *Beginners* levels of proficiency. However, during the process of examining a research question, and testing an hypothesis, the researcher will need to move even further from this, as yet theoretical and abstract level, to the concrete, practical, and real world — where the research is actually taking place! In narrowing a construct in this way, the researcher must also think about providing a definition of the "new" concept that limits this to its strictly practical, testable, observable, and measurable application in the present study and thereby also enables other researchers more adequately to replicate the investigation using the same conditions and definition applied by the "original" researcher. For this reason, the reader should not ignore such considerations, but rather look very carefully at the way a researcher has operationally defined a particular construct in the form of the variable. After all,

if we feel such a definition is deficient in some way or otherwise unacceptable, logic should warn us that the variable selected to define the pertinent construct may now no longer represent (or measure) what it is supposed to represent or measure.

The reader is encouraged critically to address these operationally-defined constructs precisely because so many of those we use in our field have acquired an acknowledged, unchallenged meaning as a consequence of their frequency in the literature. This, in turn, has given them an acquired credibility or denotation. The inherent danger is that we may then all too easily set aside any objections we may instinctively have to the accuracy with which a key construct has been operationally defined. Many of us working in the field of applied linguistics regularly refer to constructs such as *grammatical knowledge, language acquisition*, or *communicative competence* and think we all share a basic, similar understanding of what they are; however, such "shared" definitions will inevitably be influenced by personal circumstances and experience, and it follows that there might be as many operational definitions as there are people using them! For the researcher trying to discover something about a particular construct and then trying to communicate this to others working in different circumstances and with different subjects, such a potentially chaotic situation will conflict with the basic requirement that empirical research should aim to deal with precise, measurable data and be grounded in reality. Everyday communication — even among researchers! — may involve an implied acceptance of an undefined construct on both the speaker and the listener's part. On the other hand, we should expect the author to convey meanings with sufficient precision for a reader from any background to understand exactly what is being said, and in sufficient detail to permit posterior replication. After all, totally different results might be obtained if the reader carried out similar research using, say, a different definition of *bilingualism*.

An operational definition is a clear statement of how the researcher judged or identified a construct in their research through the variable. Although the question of replication and precision are important, the reader also uses these definitions to confirm that a particular construct has been defined consistently throughout the research process. As consumers of research, we need always to keep in mind any operational definition offered at this point in the paper. All such definitions will be used within a particular procedure and/or measurement in the following method or procedures section. For example, if the construct *language level* has been operationally defined by the author as the mark obtained on a multiple-choice test of French grammar, and we find this definition

— or even the test itself — in some way deficient or unacceptable, such objections will need to be recalled when we later decide on the confidence to be placed in the author's method, results, and analysis of findings.

However, these very observations also highlight some of the recurring criticisms made of this approach to definition. It has been pointed out that operational definitions may be hopelessly context-specific.[5] That is, if one reader does not like a particular definition of, say, *language proficiency* there is nothing stopping him or her providing another. Also, definitions may not always be meaningful to all readers: defining *advanced language proficiency* in terms of the number of languages someone can speak — but not write — may not be readily acceptable to many. Despite these limitations, however, it would appear the field currently feels that the clarity of communication provided by operational definitions offsets any possible drawbacks in their use.

Just how feasible it will be to answer the research questions or test the hypothesis as proposed by the researcher will largely depend on whether suitable operational definitions have been provided through the variables. In other words, operational definitions are part of the instruments with which the research question can be explored or the hypothesis tested. Assume, for example, that the researcher has told us that he or she wishes to test the hypothesis that "*L2 (Italian) students who have been taught by native-speaker Italian teachers make greater improvements than those taught by non-natives*". The onus is now on the researcher to provide us with acceptable definitions, not only of "greater improvements", but also of "native-speaker teachers" and "non-native speaker teachers". By "acceptable" I wish to indicate that the cautious reader of the research might also usefully already be reacting to the way in which the researcher is operationally defining their constructs in the variables described. As I mentioned above, this will enable us better to evaluate the information provided as a result of applying these definitions. Thus, if "greater improvements" were then operationally defined as the difference between two proficiency test scores measured at the beginning and end of a six-month teaching period, the reader might want to argue that the outcome of a test is not the "improvement" itself but only a reflection of the construct. Since other forces come into play in test-taking (and marking), we inevitably are being presented with a somewhat inadequate representation of "improvement".

---

5. See Shaughnessy, J., Zechmeister, E., and Zechmeister, J. 2000. *Research Methods in Psychology*. Boston: McGraw Hill (Chap. 1).

Our reading of this part of the paper, therefore, will need to be guided primarily by questions of accuracy. Firstly, we will need to consider if the constructs have been sufficiently narrowed down or focused to start with. Then we will need to evaluate whether the operational definition of the variable is an acceptable description of this delimited construct. In many of the papers we read, formal operational definitions may not be provided — at least not in this introductory section. In such cases, we may well need to flick ahead to the "Method and Procedures" section, as the determination of operational definitions may be based implicitly or explicitly on information given there. However, by now it should appreciated that such definitions are of prime importance to the research questions and/or hypotheses, which *are* part of this present section. It is, therefore, helpful for the reader to have an early idea of what each variable is taken to mean. Once a suitable definition has been located, we will need to consider carefully how acceptable this is in the present context. In the above example, for instance, the reader would have to decide whether the difference in two test scores is a satisfactory measurement or description of improvement in language learning. There are other contentious points here, too: within this operational definition of the construct "improvement" there is an implicit assumption that the test itself (its content and the way it was carried out) sufficiently allowed for subjects to show improvement in the first place. We might make a mental note to be on the lookout for such information in the upcoming "Method and Procedures" section. If, for example, the tests contained listening comprehension elements wherein the subjects had to hear native speakers conversing, it could be argued that the groups who had been taught regularly by such teachers had had a theoretical advantage; therefore, the tests were biased towards this group. Our confidence in the findings from such a study may be weakened since the variables themselves have not been adequately defined at the outset.

Therefore, we see that the main purpose of the operational definition in a study is to define concepts or constructs in a sufficiently detailed and concrete way to permit them to be studied, examined, measured, and replicated in future work. They are, therefore, crucial statements of intent from the researcher to the interested reader. However, defining is not an easy task for the researcher because it requires him or her to juggle concepts of inclusiveness and exclusiveness. On one hand, the researcher must aim to identify a *finite* group of observable, measurable characteristics associated with the variable. The more detail is used in this definition, the more restricted is the variable defined and the more possible it becomes to specify its exact nature. Ironically, this may

place restrictions on the generalisability of any results obtained using this definition. In other words, the author will not be able to generalise beyond this strictly-defined variable — and **generalisation** can also be a very desirable objective in empirical investigation in our field.

Let us take the case of "improvement" itself, as an L2 construct to be defined operationally. We may read that an author chooses to define this, logically, as the percentage points increase in tests of proficiency at the beginning and at the end of the experimental period. However, we could argue that this definition would not be very exclusive because some students may have improved their L2 ability but not have been able to show it in the test devised. The definition could then be broadened to include those students who, in the opinion of their teachers, were also showing improvement in their class work. However, this would admit a large element of subjective — and not easily measurable — opinion from the class teacher. Again, the operational definition would be of limited usefulness. We would be looking for a better definition that succeeds, as far as possible, in providing some observable — and, preferably, measurable — characteristic that helps us clearly differentiate between those subjects who improve in the second language from those who do not, or who do so to a lesser degree. There is no "solution" to such a situation for the author beyond their attempting to marry both needs as far as possible; however, as critical readers of this research, we will need to be aware of the consequences of such decisions as we read them.

2. Method and procedures

2.1 Subjects and materials

If the "Introduction" to the study can be said to be "what" the research is about and "why" it is being carried out, we can look to this next section of the paper to tell us "how" everything happened. Here we should find the "nuts-and-bolts" of the whole operation. The researcher, in describing what went on, will want to address two supposed conditions in the hypothetical reader: he or she can assume that the reader of the present study is an interested consumer of this research and, potentially, a person who might eventually be interested in carrying out the same or a comparable experiment in a similar context. The APA *Publication Manual* (see page 3) suggests two main aims of this section: firstly, enabling the reader to evaluate both the appropriateness of methods and the reliability and validity of results and, secondly, enabling interested readers to replicate the study. For this reason, in our appraisal of this section of a paper, we will be concerned to see how far the information the researcher provides us with theoretically enables us to repeat the study in our own contexts and also if this same information allows us to have enough confidence in the validity of the data to suggest satisfactory description and interpretation of any findings will follow.

Replication is — or should be — the basis of much of the research undertaken in our field. Foreign and second language learning goes on in many different contexts, under many vastly different conditions, involving many different kinds of teachers and teaching, themselves using much diverse material with many different learners. It follows that it will be extremely difficult ever to discover the definitive response to a research question or hypothesis found in one particular study and carried out in just one of these contexts, and which then permits us to generalise those findings to fit exactly another context of language learning. This means that much of the experimental and quasi-experimental research in our field will benefit from being replicated

in many different contexts and circumstances before any tentative generalisations can reasonably begin to be made. For a number of reasons, currently there seem to be relatively few replication studies published in second language learning research. Despite this, the crucial thing to remember here is that the critical reader — be he or she primarily interested in replicating the present study or not — needs continually to consider the degree to which the study as described could reasonably be repeated by another researcher in the same or a different context. In other words, the reader is looking for the kind and amount of information that will permit another interested researcher to carry out the study and thereby for the reader to be in a better position to assess the reliability and the generalisability of the original findings.

I have already emphasised elsewhere how the end result of our appraisal depends to a large extent on the confidence we are able to have in the study. We want to read about research that has been carried out in a manner that allows us to feel confident in the outcomes of that research. Confidence is something that is built up cumulatively throughout our critical reading. At every step, the researcher will need not only to describe what went on, but also convince us that the way he or she searched for and found answers to the question or support for the hypothesis was both reliable and valid. Particularly close study of the text will be necessary, for the reader will need to be alert to the large and diverse number of elements of the method and procedures that are susceptible to threats to its validity. Often, the way in which the researcher has devised the study itself means that certain key components of the research are not valid (see below). As a consequence, we can neither place our confidence in the use to which these components are put, nor in the findings obtained as a result of that use. The field continues to debate the relative importance of certain aspects of validity. There are so many potential threats to validity within a study that it can become virtually impossible for a researcher to control for them all.[1] Realistically, the most he or she might be able to do is show an awareness of the possible threats and plan accordingly. Part of our task as we appraise this section, therefore, will be to ascertain how far the researcher has foreseen, and dealt with, specific threats to validity, and how far any remaining questions of

1. For a fuller description of this question the reader is referred to: Underwood, B., and Shaughnessy, J. 1975. *Experimentation in Psychology*, New York: Wiley; Cook, T. and Campbell, D. 1979. *Quasi-experimentation: design and analysis issues for field settings*, Chicago: Rand McNally; Frankfort-Nachmias, C. and Nachmias, D. 1992. *Research Methods in the Social Sciences*, London: Edward Arnold.

validity are likely seriously to compromise outcomes, and thereby our confidence in the study itself.[2]

Internal validity refers to the extent to which any findings obtained are exclusively the result of the variables being studied here or are potentially affected by other factors that are not part of the original relationship studied. These factors may derive from any number of aspects related to the study, but mostly arise from the research design and/or the procedures used (see below). Just as there are factors which affect internal validity, there are others which compromise **external validity**. However, it is important to bear in mind that if a study does not have internal validity, it cannot have external validity. This is because external validity allows the researcher to generalise beyond the present data to other contexts. Logically, if the outcomes described are already seen to be jeopardised by questions of internal validity, we can have little confidence when we attempt to infer from them to other contexts beyond this present study. However, this does not mean that one kind of validity is a direct consequence of another. Much of the research we read will be well controlled and designed, so that there are no serious objections to internal validity. In this case, the researcher (and the reader) will be able to have reasonable confidence in the *description* of any outcomes obtained.

However, trade-offs will inevitably need to have been made by the researcher: the very control or restrictions placed on components (which was needed to achieve such internal validity) may, in turn, mean that the study becomes so far distanced from the reality of other second language learning contexts as to have no external validity (cf., the argument in the previous chapter concerning the accuracy of operational definitions). In appraising the many quasi-experimental studies with **intact groups** in our field, consideration of threats to internal validity will be paramount. If any major threats are found in the setting up of the study, the acceptability of any description of data may be seriously compromised. Where **true experiments** have been undertaken, we will need to pay particular attention to the way the researcher specifically overcomes any relevant threats to external validity presented by the subjects and procedures used, so that we are then in a position to evaluate generalisations or inferences based on findings (see below).

---

2. More recently, Wampold et al., expanded the concept of validity to include the theoretical context of research and also potential threats to what they refer to as "hypothesis validity". (Wampold, B., Davis, B., and Good, R. 1990. Hypothesis validity of clinical research. *Journal of Consulting and Clinical Psychology,* 58, 360–367.

Subjects — Replication

What basic identification data are provided about the subjects, and are these data sufficient to permit replication?

The presentation of the section in the paper known as the "Method" will vary according to the type of study being read. Typically, however, a section will be provided wherein the source of the data to be collected is described. In the studies we are concerned with in this book, we will often find language learners and their characteristics as the main source of data. However, it is as well to remember that data can come from many sources, such as pieces of text, specially-selected or randomly-selected words from a corpus, or utterances from one case-study subject. Similar considerations to those that follow will need to be made in the appraisal of such sources. So, for example, if the source of data turned out to be a transcript of speeches made at an L2 teaching congress, and we were interested in using the same data in our own replication, that data would need to be readily-available for consultation. Similarly, if data were anonymous responses to a test, these would usefully be supplied by the researcher in an appendix or available on request.

The kind of information about learners and their characteristics which the interested reader would need to replicate the study is what we might call basic identification data. Firstly, who were the subjects, and how many were involved both prior to and during the research itself? As regards the "who" were involved, the researcher would want to supply sufficient information to make replication feasible. There are no strict rules about what information should be conveyed; it inevitably depends on the study itself and our own knowledge of the field. For example, gender might need to be known, as it has often been shown to be an important variable in our field. Age has been similarly shown to be a central factor in some aspects of language-learning. The subjects' previous academic (language-learning and other) experience may play an important part in outcomes. Doubtless, we can imagine how these and other basic factors such as nationality, the native language of subjects, current course and place of study, or their attested level of L2 proficiency may all be of central importance in certain studies of second language learning. However, other, more subtle, characteristics may also impinge on the way subjects act and react in a second language learning situation. Knowledge of another (i.e., third) foreign language can often be a help or a burden. Moreover, many studies have indicated that ability and desire to learn a second language may not be enough when the socio-economic background of these subjects does not permit them to compete

on equal terms with other students who are more easily able to access such resources as internet, exchange visits, or satellite television.

The researcher will also have decided — according to his or her reading of the circumstances — whether individual, group, or average data are given in such basic identification categories. Much will depend on the way results are to be presented and what is to be read into these findings. Just what, and how much, detail is supplied will depend on the researcher's judgement — which the reader will then need to appraise. Our understanding of what happened is not necessarily enhanced by describing the subjects and the context in minute detail. Indeed, there will be many occasions when superfluous detail about the subjects and the context of the study only succeeds in confusing matters. In practice, the reader should read the basic identification data provided in this section, think about what he or she has been told so far in the study, and make note of any apparent deficiencies. Once again, however, whether any highly relevant information about subjects is lacking here may only become apparent as we read on in the paper. Hence, I emphasise the renewed need for the critical reader constantly to be aware and to note down anything he or she will want to follow up as a result of subsequent findings.

a. What are your initial reactions to the numbers involved or any grouping envisaged?

b. Do these groups reflect the original pre-group sample in terms of their basic characteristics and is any justification provided for the eventual group size?

Although we will probably not yet know the kind of research design envisaged or the analysis to be carried out on the data, the kind of conclusions we can draw from findings will often depend — amongst other considerations — on the numbers of subjects involved and whether these numbers remained constant during the period of data-gathering (see below). Here it would be as well to pause and make an initial response to the logic of the method so far described. In other words, given what we know about the aims of the study gleaned from previous sections, do the numbers of subjects involved appear to suggest that useful data will be collected? For example, in a study where the research question aimed to gather information about typical classroom L2 learning strategies, one might wonder whether three groups described as each consisting of four subjects might provide sufficient data to be adequately studied. Finally, another useful awareness-raising exercise during our reading of this section is to think about what data might *not* be obtained as a result of

such numbers of subjects. For example, while large numbers of participants from several classes may well provide the researcher with considerable support for any collective responses observed, we will learn nothing about the individual processes involved in forming such group reactions. This is not a fault of the research, of course, nor the researcher, who was interested in other things. But such reflection does have the advantage of helping us set our own minds thinking about other explanations for any outcomes described and, perhaps, helps us form ideas for our own future research agendas.

Subjects — Internal Validity

Crucial factors such as which subjects are selected, and how, can affect the internal validity of a study and will need to be appraised carefully. We would be interested to see how the researcher planned to control any sources of possible bias in the subjects. The most obvious component of subject selection for the researcher to have validated is that all the subjects used actually match any description made of the group — in other words, that any groups eventually to be compared are on equal terms to start off with. Thus, if the researcher has decided that only female students, of between 15 to 20 years old and with intermediate level English are to participate in the study, the researcher would want to have taken the necessary steps to ensure that all subjects actually fit that profile. Bias might unwittingly be built into studies where the original sample is subsequently divided into groups. If we read of an initial sample of subjects who — according to the abstract — were then divided into four groups, we might want to consider whether these groups still reflected the initial balance in the sample in terms of gender, proficiency level, age, and so on. The reader needs to read about the initial sample characteristics and — if the objective so demanded — understand how far the sub-groups formed still reflected these characteristics, and then consider the extent to which any deviation might seriously affect any outcomes.[3] Imagine, for example, that a researcher tells us that he or she started off with an original sample of 50 subjects, 35 of whom were in an advanced-level class, the remaining 15 lower-intermediate level. If we then understand from the abstract that one **control group** (sometimes also referred to as a "comparison group") and one experimental group were

3. A more detailed discussion of sampling techniques can be found in McCready, W. Applying sampling procedures, in Leong, F., and Austin, J. (eds.) 1996. *The Psychology Research Handbook.* Thousand Oaks: Sage (pp. 98–112).

subsequently formed with 25 subjects each, we would need to be alert to the possible consequences of such implicit imbalance.

A researcher may choose groups on the basis of the highest and lowest scores on a **pre-test**, thereby trying to achieve one group of "weak" subjects and another of "strong" subjects. A similar situation can be brought about by the common practice of selecting a group "in need of special treatment", such as a class of "poor language learners" who the researcher guesses will (and do) perform poorly on the pre-test. Statistical analyses have shown that chance factors very often play a key role in high or low scores and that these factors are unlikely to reappear on a second or **post-test**. Therefore, what is likely to happen with our "chosen" groups on a second or post-test is that more of the original "higher" and "lower" scorers get results which move more towards the average (known as "regression to the mean"). In other words, a difference would have been noted between the groups on this measure, even if there had been no treatment in between the two tests. Any findings resulting from such a procedure would then be of doubtful validity.

Selection bias of this kind can an acute problem in many of the quasi-experimental studies in our field because we are so often obliged to carry out research with intact classes (i.e., classes to which the students have been assigned prior to the study itself). There may be an implicit assumption that such classes already reflect random assignment of groups because they are somehow a suitable "mix" of students. The study then proceeds with analyses and research designs that are suitable only for truly randomised groups of subjects. However, even where students are placed in different classes at the same level of proficiency, the way they are assigned to those classes is usually not random and corresponds to other considerations, such as scores on tests, alphabetical order, or timetable considerations. Students might even have self-selected a class on the basis of the teacher or the time of the class. Likewise, teachers might even select who participates in which class because they wish to have more homogeneity in the group they are teaching. Working with such intact groups does not preclude studies involving control, experimental groups, and a particular treatment. What can be done in these cases is to use findings to present support for some kind of effect of, or relationship between, the variables described. Such findings, however, will need to be read and interpreted with care and few, if any, conclusions about similar results in other contexts could be drawn without replicate studies with many other similar classes.

This said, basic designs using intact groups can be improved upon (see below, "Procedures"), and our confidence in the validity of the relationship data

can be increased by observing how a researcher sought to control or reduce any potential selection bias. For example, it is relatively easy to flick a coin and randomly assign the control and experimental groups, although random assignment of these intact groups to such groups will not be sufficient to balance any systematic differences among the intact groups. The researcher might also decide to restrict the original **population** from which the classes are drawn (e.g., only beginners' level of language proficiency) and thereby keep in check any selection bias introduced by using the first classes that come to hand. The problem that the reader will still need to evaluate is how far this restriction limits any conclusions drawn only to this specially-limited group of individuals.[4]

The size of subject groups will also be a cause for concern in many studies. While there are few fixed rules for ideal numbers of participants required for the majority of analyses used in second language learning research, recommendations do exist and will need to be considered when appraising research design.[5] We may be faced with a description of the numbers in each group, but little more. However, the determination of group size should ideally respond to some principle, particularly in the case of studies that will later involve some descriptive or inferential statistical analysis of the data. In this case, for example, a researcher may decide to increase the size of a group of subjects with an eye to increasing the **power** of the study (see below). Thus, **non-parametric statistical tests** are often used in our field because they carry fewer crucial assumptions about subject selection than **parametric tests**. One recommended way of increasing the power of such tests (and, thereby, the confidence to be placed in

4. Many of these aspects of selection and assignment to groups reported on assume that the researcher has adequately respected the relevant ethical standards in the research. The failure to follow such codes needs to be assessed by the reader for it can undermine the study itself and, ultimately, the entire scientific process. The *American Psychological Association* (www.apa.org) has produced a list of "Ethical Standards" which cover both the preparation and the reporting of studies: Ethical principles of psychologists and code of conduct. *American Psychologist*, 1992, 47, 1597–1611. Readers are also directed to Oleson, K., and Arkin, R. Reviewing and evaluating a research article, in Leong, F., and Austin, J. (eds.) 1996. *The Psychology Research Handbook.* Thousand Oaks: Sage (pp. 40–55) for further ideas on appraising ethical questions in research papers.

5. See, for example, Kish, L. 1965. *Survey sampling.* New York: John Wiley; Mehta, C., and Patel, N. 1995. *SPSS Exact Tests 6. 1 For Windows.* Chicago: SPSS; Keppel, G. 1991. *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: John Wiley.

the results) is by increasing the number of subjects.[6]

The fewer subjects used, the more likely it is that sufficient bias will be introduced into the sample potentially to distort the data provided. The reason for this is that each member in a small group of subjects will perforce bring about a greater effect on the overall group performance than would have been the case in a larger group. In our appraisal of this component, we will want to weigh up the requirements of any group statistical analysis foreseen, since excessively low numbers of subjects in each group may well have the effect of distorting key measurements such as the mean or **median** and central measurements of **variability**, such as the **variance** or standard deviation. On the other hand, group size will be of lesser importance where the research is basically concerned with individual variability, such as in the case of describing the way in which a select group of L2 students use certain communicative functions, and where inferential statistical analysis is not to be an issue.

Can you see any potential threats to internal validity of the data from attrition, history, or maturation factors?

We will also need to think about the potential for attrition (also known as "mortality") in the subject sample. Once again, we are referring to possibly troublesome variation within the sample selected, but this time the problem is that subjects drop out of the group during the study. This is a particular problem in **longitudinal studies** since the researcher obviously does not know in advance who will not continue, and the absence of certain members from groups may seriously affect outcomes and/or the interpretation of these. Those subjects that remain in the group may present different characteristics from those that have left. Furthermore, if all the dropouts come from one group, this latter may then present data which have been directly affected by this imbalance, rather than as a direct result of the variables being observed. Consequently, any comparison between groups will be seriously undermined and the validity of any comparative data brought into question. The reader will need to be alert — particularly in longitudinal studies — to the information given about dropout rates and consider whether sufficient explanation has been given for subjects' withdrawal and whether this has been adequately accounted for in any posterior analysis.

---

6. Further advice and other perspectives on the threats to validity posed by subject selection can be found in Shaughnessy, J., Zechmeister, E., and Zechmeister, J. 2000. *Research Methods in Psychology*. Boston: McGraw Hill.

History and maturation are two threats to validity which often need to be considered, in particular when we are reading about studies carried out over relatively long periods. Although papers may give the impression that subjects have temporarily dedicated their academic lives to the study, many other things could be happening to the subjects — or their L2 language input — while this is going on! Such history factors are often to be expected in longitudinal studies, particularly when research is taking place in contexts where the target language is already spoken as a second language outside the "laboratory". We might imagine, for example, that subjects are communicating to some extent in that language in their everyday lives and hearing the target language around them. Such uncontrolled input will doubtless have some sort of effect on any data gathered about learning within the L2 classroom. This is less of a problem in foreign language learning contexts but, even so, we might reasonably expect the researcher at least to check up on (even if he or she is unable to control) any possible sources of interference or distortion in the data gathered. One might conceive of a situation, for example, wherein a researcher is testing the effect of a new set of reading and writing materials and is unaware that, concurrent with the research, a number of subjects are in regular L2 correspondence with pen-friends or take part in frequent computer chats or e-mail exchanges in that language.

Admittedly, for the reader, we are likely to remain unaware of these factors, but there are language-learning contexts that are used for research in which we can hazard a good guess about what outside influences might potentially be seriously affecting the study. Either way, it would be a sound strategy on the part of the researcher to anticipate such observations and provide us (as far as he or she can) with details about how any potential influence was sought to be controlled. One of a number of ways around the problem would be to provide a control group who are thought to be experiencing the same or similar history factors as those in the experimental group. It should go without saying that the researcher ought to have ensured that any experimental and control groups experience the same history events as part of the research design: both groups to be compared should be experiencing the same teaching, learning, and/or testing conditions, apart from the variables currently being examined (see below, "Procedures"). The effect of the teacher, for example, might be controlled by having the same teacher teach both groups. Classroom conditions should be similar: if the reader feels, for in-

stance, that the control group are at a disadvantage because, unlike the experimental group, they had their class in the late evening, this history factor might be discussed as a possible threat to internal validity.

Maturation is also related to time and affects internal validity when we suspect growth, change, or development in the subjects related to the treatment studied in the research. Once again, this is a particular problem in studies over longer periods of time. We know that young people of a certain age, for example, experience considerably more cognitive development over similar periods of time than adults. Thus, we might not be surprised to find changes in L2 learning ability in such subjects caused by factors other than those currently under study. The developing ability to use short-term memory to greater effect may be a particular feature of a child's cognitive ability that could affect results. In addition to actually getting older, maturation might also refer to coexisting elements of long-term classroom experience, including tiredness and boredom. Likewise, care would also need to be taken when comparison is made between groups comprising different age-groups. Adults, with more maturity and greater experience in the classroom, might be hypothesised as better able to handle certain methodologies and specific evaluation measures than younger subjects. Once again the reader will be looking for some acknowledgment that maturation could have occurred (it is difficult to control, of course) and suggestions as to how some of these interfering experiences were confronted in practice.

Subjects — External Validity

> **What information is presented concerning the way subjects were selected and/or group membership assigned? What do you see as the consequences of this as regards eventual generalisation of findings?**

As far as our appraisal of true experimental (as opposed to quasi-experimental) research is concerned, the key point to look for will once again be in the way subjects are selected for the study. The question we need to ask when generalisations are intended is whether the subjects and the context in the study are representative of the subjects and context to which the researcher apparently wants to apply the findings. Thus, for example, we would rightly question whether findings about the abilities of a group of teenage beginner L2 language learners can be generalised to adult beginners. Similarly, it is of doubtful external validity to generalise improvements in L2 proficiency after using an

experimental language laboratory program with university students to groups of parents learning the L2 at night-school, where context, subject characteristics, interests, and motivation may be very different.[7]

However, before such generalisations are ventured (usually in the "Conclusions" section of the paper), there are a number of elements to look for in the way the researcher goes about forming groups. True experimental designs may look to achieve a representative sample of the population by using **random selection**, such as allocating numbers to subjects and then choosing them out of a hat. One of the unfortunate consequences of such a method is that the researcher might then end up with rather more male than female, more proficient than less proficient, or more immigrant than foreign subjects in the final sample. In other words, we should initially regard such evidence of randomisation as rather simple and potentially of limited use for eventual inferential conclusions. More confidence should be placed in grouping that results from previous **stratified randomisation** using selection within specified strata, such sex, age, or region. This suggests a more precise search for data on the part of the researcher: specific previously-identified characteristics are to be represented in the randomised sample and that sample reflects the true characteristics of the population to which generalisation will be made. In such cases, the reader might benefit from being told what these characteristics actually were and on what basis these were identified as being important in the present study.

We would then need to read about how subjects, once selected, were subsequently assigned to any groups. As mentioned above, and in the interests of obtaining sound internal and external validity, group membership might also be assigned randomly. At this point we might also need to know whether subjects are then to be **matched**, with an eye to subsequent statistical procedures, and what the basis is for this matching.

The method used to select subjects and assign these to groups will be crucial for any posterior analysis and discussion of data, and the reader will need to appraise how far the main threats to external validity have been met. On the one hand, the researcher may decide to use statistics only to describe data and to

7. It has also been argued that in some true experiments external validity might *NOT* need to be established since, if the objective is to test a particular hypothesis drawn from a psychological theory, an experiment might have been undertaken to see whether subjects can be encouraged to react in a particular way. Whether they do so outside the "laboratory" may not be of interest. (Mook, D. 1983. In defense of external invalidity. *American Psychologist,* 38, 379–387.

assure us of the validity of that description. For the description to be perceived by us as accurate, we would need to confirm that all major threats to the internal validity of the data have been met. On the other hand, the same statistical procedures might be foreseen by the researcher to permit generalisation from these data. The critical reader would then want to be sure, at least, that a detailed description of the target population is provided and, secondly, that random selection has been appropriately used. Finally, all other major threats to internal and external validity (i.e., including those in sections below) should be seen to have been adequately met.

Materials — Replication

> **Has any material or instrument of testing/measurement been satisfactorily described and/or samples provided? Where appropriate, has its development/design and scoring been adequately discussed?**

Once again, the key to replication is the detail provided by the researcher. Typically, limited space is assigned to published papers in our field, so there will be few chances of our being provided with complete copies of materials or instruments of testing used together with the published study. Given these constraints, the reader would still benefit from seeing some key examples of what subjects saw and used, perhaps together with more detailed samples in the appendix, and/or more information as to where the complete materials or instruments can be accessed, if these are not already well-known in the field. Detail includes not only the description of the data-gathering instrument itself[8] but also, where appropriate, thorough informative discussion of its design development, items, scales, and scoring. Therefore, for example, if a questionnaire to subjects has been created or an existing one redesigned, it would be useful to be told in general terms how and why the researcher set about designing or redesigning the questions, and what kind of information was sought through these actions. Likewise, learning about how individual responses were weighted as scores will then put the reader in a better position to appraise the way results were obtained and what they mean. In such a way, we can begin to make up our minds whether the questionnaire itself was a valid instrument as such (see below, internal validity) and/or use the information for replicating or

8. "Data-gathering instruments" include tests, observations, or any other formal means of obtaining data. References to "tests" may include any of these means of data collection.

adapting the questionnaire using the same criteria as in the present paper, but with our own subjects.

Similar considerations need to be made as regards any technology used in the study. Increasingly, one comes across studies wherein use has been made of such things as overhead projectors, tape-recorders, video cameras, or computers. Like other materials, these are all important vehicles through which treatments may be administered. Thus, the researcher must be concerned to describe this apparatus and the use to which it was put in enough detail to permit a similar experiment to be carried out again.

Materials — Internal and External Validity

For any instrument of testing or measurement used (including observation), what evidence of reliability was given and how acceptable is this evidence?

To be valid in a specific data-gathering context, materials used must truly be designed to reflect what they are supposed to describe. It is vital, therefore, that whatever procedure is used for collecting the data has both acceptable validity and reliability. Both factors will need to be addressed by the researcher and critically examined by the reader. For example, if we read that a researcher intends to use an existing instrument of testing or measurement, we would expect to read of, and assess, the established reliability and validity of items and scales on this instrument. Moreover, we expect soon to read of the way data were analysed in the study: for such an analysis to be selected and carried out appropriately, the researcher is implicitly accepting that the data are both reliable *and* valid. As we shall see, each statistical test has specific assumptions related to it, but all begin with the one assumption that the data being fed in to any formula respond to this basic requirement.[9]

Reliability may be judged informally by the reader as the extent to which we believe (or we are told) the data-gathering instrument might produce consistent and accurate results when it is given under similar conditions elsewhere. Reliability may be formally reflected in agreement between observers or scorers

---

9. Judging reliability need not always be so rigid. Where the research interest is in *individual* scores and performance, unreliable instruments will be a serious flaw. Yet if the researcher is interested in average results for posterior group comparison, some deficiency in reliability might be tolerable (see further discussion of the effects of unreliability on different samples in Christensen, L. 1980. *Experimental methodology*. Boston: Allyn and Bacon).

of a particular phenomenon or, in the case of instruments of testing, in consistency coefficients (see below). In the absence of, or in addition to, such official verification there are a number of less formal factors that we might want to consider in appraising the reliability of any material in a study. Firstly, we might wonder whether the way the instrument is constructed looks as if it promises accurate data-collection: perhaps the samples or test questions themselves seem to be ambiguous when we read them and/or we think they could not be answered accurately or honestly by these students. To reassure us, we might usefully read what steps were taken to pilot the instrument and report back on what items proved reliable and which did not. Similarly, it would be interesting to be told with whom this field testing was undertaken (ideally, a very similar sample to the final target group), and what, if any, provision was made to gather their opinions on the instrument.

Secondly, we need to ponder whether the conditions in which subjects use the instrument might reflect on the reliability of data gathered from its use. Perhaps we might think subjects could easily get fatigued because it is so long and/or complicated, or perhaps there is an over strenuous time-limit on responses that we imagine might have induced unnecessary tension in the respondents. Thirdly, there is an argument that subjects need to have some initial familiarity with the specific data-collection instrument to produce reliable results: we might imagine how initially difficult it can be for subjects who have never had to face a questionnaire with **Likert-scale** responses (i.e., 1–*strongly agree*, 2–*partially agree*, 3–*agree*, etc.) to decide on their answers. Data obtained in this way may be compromised, not only by the previous effect of a variable, but also the subjects' lack of familiarity with such a questionnaire format.

As we saw in the treatment of threats to validity, it is rarely possible to meet all threats of reliability. Indeed, it might not always be a good idea to strive for such an aim: a more stringent control on reliability might result in a less desirable trade-off with realism. However, once again, our confidence in the readings made of data can be strengthened by seeing the ways in which a researcher has sought to recognise and meet some of these threats. There are a small number of ways in which a researcher can estimate more formally the reliability of any test instrument. These are useful reporting devices to increase reader confidence in the instrument being used. The reader might look out for

the reporting of **reliability coefficients**, with 1.0 indicating a perfectly reliable test.[10] Amongst the many we may come across in this section of the paper are internal consistency reliability coefficients derived from tests such as the **Kuder-Richardson** or **split-half** tests (which calculate the coefficient based on dividing the test data into two similar parts) or test-retest reliability coefficients (where the reliability is tested over time and as a **correlation** between test and retest scores). We might also come across references in second language learning studies to "**inter-rater/observer reliability**", wherein two or more raters' scores using a particular instrument are compared in an effort to establish reliability for the instrument of testing and the raters themselves. Even when a researcher is using an instrument that has already been shown (and suitably reported) to be reliable and valid in previous studies, the researcher (and the reader) would want to be assured of its reliability and validity in the present study. One previously acceptable coefficient in one context is no guarantee of future acceptability in another.[11]

Whatever the route chosen to obtain the reliability coefficient of an instrument of testing, we should consider both the measure and the coefficient obtained with caution: the figure tells us just how reliable a major component of the study has been shown to be, and any low reliability reported will need to be adequately addressed by the researcher. Our confidence in the outcomes should again be suitably bolstered or undermined as a result.

> **For any instrument of testing or measurement used, what evidence of validity was given and how acceptable is this evidence? If none is given, what do you consider to be possible threats to validity here?**

Once reliability has been acceptably established, the reader would want to see how far the researcher has then considered the validity of the instrument being used. In other words, to what extent is the researcher convinced that the

---

10. Estimates of what constitutes an acceptable coefficient vary and much depends on the purpose of the instrument in question. More detailed information on suggested acceptable coefficient levels in different types of research can be found in Carmines, E., and Zeller, R. 1979. *Reliability and validity assessment.* Beverly Hills, CA: Sage and in Pedhazur, E., and Schmelkin, L. 1991. *Measurement, design, and analysis: an integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

11. Carmines and Zeller (1979, op. cit.) provide a more detailed description of other methods of reliability testing, including Re-test, Alternate-form, and Internal consistency methods.

instrument succeeds in obtaining data about what has already been operationally defined as the variable in question, and not something else? There are different kinds of validity, but four are fundamental for work in our field. **Face validity** will be particularly important in our appraisal of data-collection, as it refers to the researcher's (and, in our present context, the reader's) subjective appraisal of what the instrument is measuring. We should ideally be able to read the items, or a representative sample of these, and conclude whether they do, or do not, *appear* to be measuring the construct the researcher intended or the variable to be measured. It is, therefore, an informal judgement (unlike **Content validity**, see below) that should be made. Such intuitive judgements about validity can be extended also to other questions: we might ask ourselves whether the length of the instrument appears to be appropriate for the proposed use or whether it seems to require too much or too little time to complete in order to obtain adequate data. Similarly, we might want to know whether the wording in, or readability of, the instrument (particularly if it was presented in the target language) could in any way have inhibited responses from these subjects.

Content validity is a more subjective and formal evaluation of the instrument. It describes how far the contents represent a demonstrative sub-set of what the whole instrument is supposed to evaluate. It involves clearly defining the construct being considered, selecting a sub-set of this construct for the instrument, and finally operationalising these as items in the instrument. Obviously, operational definitions already established will be important here: if a researcher purports to use the materials to produce data about proficiency in L2 grammar, we would need to decide (or be told) whether a representative and complete sample of "L2 grammar" had been chosen for the instrument of testing. If we see from a sample of the test, or a report of its contents, that such a test comprises only multiple-choice type questions on verb forms, we would be right to doubt the content validity of the instrument for "L2 grammar". Since there is no specific descriptive or inferential statistical analysis available for content validity, our appraisal of this element will need to be based on the accuracy of previous operational definitions, on samples, and on common sense. The researcher, however, in attempting to establish content validity should previously be seen to have attempted other formal means: for example, the researcher could have previously submitted the questions to a group of professionals and asked them to rate the representativeness and comprehensiveness of the content. However, the reader would then need to have enough information about this particular group to accept that they were sufficiently unbiased themselves to provide acceptable judgements on this content!

Predictive validity, as the name suggests, describes how valid the instrument is for the prediction of future outcomes. So, we are being asked to accept that a poor result on our L2 grammar test will be a good predictor of equally poor outcomes in other non-test contexts. The assumption here is that the instrument is robust and valid enough to predict results beyond its present confines. Logically, validation would need to include some kind of **correlation coefficient** based on the original instrument and the non-test context.

Construct validity has been an underlying feature of much of our appraisal work so far and is a particularly difficult construct to validate, given that there are so many concepts in our field which lack obvious methods of measurement. It is a critical validity to establish for any quantification process because a construct is the perception the researcher has of what he or she is actually quantifying. Indeed some writers in the field go so far as to suggest that the other types of validity actually revolve around construct validity or are even included in construct validity.[12] It is established by demonstrating that a particular instrument succeeds in measuring a specific construct. A number of tests or processes are used with the instrument concerned to establish construct validity, and it would be the accumulated outcomes of these which would succeed in convincing us (or not) of the validity.[13] If a researcher claimed to be measuring the construct "advanced level L2 reading comprehension" with their instrument, he or she could find one group of subjects who are L2 intermediate level and one group who are advanced L2 students and try to validate the construct validity of the instrument by showing that the advanced students scored higher on the test than the intermediate students. However, the result from one such test and one such sample should not succeed in convincing the critical reader of the construct validity. A number of other variables might interfere with the soundness of results from one test. Much would depend on the use the researcher then wishes to make of construct validity in their study, but we should remain sceptical of the establishment of such validity based on limited testing of the instrument in too exclusive a context. Furthermore, other scholars have pointed out that unforeseen results from these tests (i.e, apparently demonstrating the invalidity of the instrument) might actually be providing evidence of the erroneous *theoretical*

---

12. See Carmines and Zeller (1979) and Pedhazur and Schmelkin (1991): footnote 10.

13. See Brown, J. 1991. *Understanding Research in Second Language Learning.* Cambridge: Cambridge University Press, p. 103.

viewpoint previously held about a particular construct rather than showing the invalid *measurement* of that construct in this instance. In other words, establishing construct validity for a test or other quantification instrument does not necessarily question the concept of the original theoretical construct itself, where we may or may not agree with the author.[14]

Apart from those mentioned in passing above, there are a number of other specific threats to validity of materials that the reader will do well to consider as he or she appraises this section. In most cases, these are questions of logic, but we may well find that they are not sufficiently addressed in the text and that more specific information is now required. Firstly, in the case of materials used in previous studies, it is good practice to read carefully any description of the source of such material and subsequent adaptations of this with a view to determining how valid this present application of the material may be. So, for example, we may be told that English-as-a-foreign-language subjects' compositions are to be evaluated using an un-adapted version of a marking scheme designed for use in English-as-a-second-language contexts. We would need to decide (from prior knowledge of the instruments involved or detailed examples of the same) how far the scheme is valid for these present purposes or groups, given the EFL group's different learning context, needs, and objectives.

Similarly, the content of any material used will need to be appraised for its appropriateness to the present subjects. What needs to be considered here is whether these subjects are able to show their real competence given the design of the procedure. We might read of a listening comprehension passage that is being used as a post-treatment test of an innovative classroom methodology; however, as we read through the details, we note that the subject-matter of one of the passages is specifically concerned with economic science. As a result, the vocabulary used could be of a rather too specific nature for these subjects. Arguably, therefore, the content is no longer valid for this population as a test of listening comprehension.

How have the main independent and dependent variables been realised within the method itself, and how satisfactory do you find this?

Once the materials to be used in the study have been appropriately described and justified, we can now attempt to link up what we know so far of the "how" of the study with what we have already read of the "what" and "why". It is, once again,

14. See Scholfield, P. 1995. *Quantifying Language.* Clevedon: Multilingual Matters.

another place in our appraisal of the paper where we are advised to stop reading for a few minutes, digest what we have been told so far, and think about the consequences. In the interests of space, published articles rarely dedicate a specific section to the way the main variables are realised in the method itself. However, for the reader, thinking about such concerns is invaluable to help us follow the logic of the study. Specifically, the reader should be in a position to see how the measurement and observation of the variables previously described in the problem statement and research question/hypothesis have now been directly related to the materials used. If the relationship is not made evident by the researcher, the reader should try to confirm the appropriateness of the assignment of variables in the method. Once again, we should be primarily concerned with consistency between sections of the paper, based on what we have read and reacted to in previous sections. Check particularly to see if all this coincides with what had been expected from the description provided in the research question or hypothesis, and confirm if any doubts felt at that point have now been explained satisfactorily or increased as a result of our reading of the method so far. We might also usefully confirm how any levels of the variable predicted in the previous section have been realised in the method.

2.2 Procedures

Now that we have read about the subjects and instruments involved, we now need to be told how the two interacted. In other words, we will now appraise what is arguably the most critical component of the "Method and Procedures" section. We now expect to be told what happened to the subjects from the beginning to the end of sessions in which they were involved. Once again, possible replication is facilitated if what the subjects did and what happened to them is recounted in step-by-step chronological order. This will also give the reader the chance to consider how the context of the investigation might have been changing for these subjects throughout the study. Since many studies in the field involve the assignment of subjects to groups, it is also worthwhile trying to follow how each group was treated separately throughout the procedure. Our understanding of what went on can be aided if we are provided with separate descriptions of each group procedure or a summary description of all the groups involved followed by detail of the distinguishing features of each.

What is your appraisal of the timing of events, and is this information sufficient to permit replication?

Apart from looking for the kind of detail in the description that permits adequate replication of the study, the critical reader will also need to be attending to possible further threats to internal and external validity of the data. Firstly, we might usefully think about the consequences of the timing of events as they are described: for example, is it felt that the time allowed for different procedures was sufficient given the nature of the experiment and/or the L2 ability of the subjects? Or perhaps we read that the period of time elapsing between the pre-test and the post-test is short and, as a consequence, we feel that the threat exists that subjects remembered items from the pre-test which then affected their results on the same post-test content. Likewise, we might want to consider the period of time between any treatment and a post-test to evaluate how hypothesised changes were monitored, and how much confidence to place in these observations. For example, we might well be wary of any reported improvements in L2 performance shortly after the treatment period: it would be quite usual to discover a sudden and transitory increase in subjects' motivation, and thereby perhaps language performance, as a result of recent innovatory methods in the classroom during the treatment. Lasting changes in language performance, however, are usually the result of much more subtle processes that require closer and more prolonged observation of subjects. We often read that studies are set within a specific time period, "during an academic year", "for two months in the spring term", or "over a semester". What needs to be borne in mind, however, as we consider these events is that — unless previous studies have indicated specific time periods — it is often impossible for the researcher to suggest beforehand a fixed time within which improvement ought be noted as a result of the treatment. Certain research designs can help increase our confidence in any claims made by the researcher in this regard (see below), but there are no specific rules available to follow. The researcher (and the reader) will need to consider the evidence for change in the light of the time periods within and after the treatment, the context of the data-collection, and the appropriateness of the data-collecting instrument itself (see below).

Are there any potential threats to internal validity as a result of test or practice effect?

The period of time elapsing between tests is not the only potential threat to internal validity arising from the way data-gathering is carried out. Being exposed to

some kind of test often has an effect on any posterior test, particularly — but not only — if it is the same one. Research into short- and long-term memory has shown that while memory for the *form* of the questions may fade relatively rapidly, memory for content is more long-lasting. There is often a familiarity with the test procedure obtained as well as a knowledge of one's mistakes and, indirectly, an acquired insight into what the researcher is testing in the instrument in question. Having taken the pre-test, the assiduous subject in the control or experimental group might have been made aware of their weaknesses, want to try to improve their performance, go off and bone up on what will in fact constitute the treatment, and return to obtain a better performance on the post-test as a result of these efforts!

Performance in the post-test might, therefore, be enhanced as a result of such test or practice effect, rather than the treatment itself. Another, more subtle threat we need to be alert to as a result of pre-test procedures is when the latter aim to collect data about students' affective characteristics, such as attitudes or feelings. The aim of such pre-tests, of course, is precisely to probe deeply-held opinions or postures and assess them. Unfortunately, what may also happen as a result of this test is that subjects end up becoming more aware of their own attitudes and thereby are alerted to how they may develop as a result of the treatment. If the post-test examines the same attitudes, the responses obtained may no longer reflect the effect of the treatment so much as the effect of each subject's reflections on his or her attitudes as exposed through the pre-test. We would also need to be particularly attentive to the time periods between each data **sampling**. In longitudinal and/or **repeated-measures** studies, where the same subjects might be tested a number of times on the same measure, we should remember that subjects are also simultaneously practising that task, and we must expect to see change in these subjects — even if a control group is involved and all the testing is done under the same conditions. Thus, we might see experimental subjects improve because of the treatment they receive, and control group subjects get better (although to a lesser extent) because of the practice they are getting anyway. Conversely, the control subjects might get worse because of factors such as boredom, frustration, or tiredness.

As with any potential threat to validity, and with a view to increasing our confidence in the procedures used, the reader might be interested to see the extent to which the researcher acknowledges (rather than ignores) the existence of such threats and attempts to counter or allay them. For example, there are research designs (see below) that meet some of these threats by avoiding the pre-test altogether or attempting to get hold of the required data indirectly.

Other researchers might opt for the more complicated route of having a number of equivalent pre-tests to use on the different groups (known as counter-balancing).

What is your assessment of any instructions given the subjects?

Another important aspect to be appraised in reading about "what happened to the subjects" concerns any key instructions they received. In the interests of replication, of course, those instructions should ideally be reported verbatim or summarised accurately. In the interests of assessing the potential threat to internal validity, however, any directions given subjects would need to have been appropriately checked themselves, normally as part of a previous piloting procedure. An oft overlooked aspect of internal validity, this can be a particularly acute problem when instructions are given to language learners, even more so when those instructions are given in the target language itself! Instructions — in whatever language or medium they are presented — need to be fully and correctly understood by all the subjects for subsequent outcomes to be valid. Ideally, we should be able to read through a transcript of what was said to subjects or what they read before data-gathering commenced. Our attention could be drawn to potentially confusing or ambiguous elements here, particularly when created as a result of lengthy directions. Similarly, we might include here for appraisal any instructions received by the subjects for revealing personal details. Imagine, for example, that subjects have been told (as we would have hoped) that all data would be treated in confidence and that personal identifying details were not relevant to the questionnaire they are about to answer. If then, we — and the subjects — see that they are asked for their names, class designations, or other identifying information, we might suggest that certain data are potentially compromised since subjects may now be more wary about the responses they give. Both this and the following (reactivity) issue again touch on the question of the ethical standards used in the research and readers would do well to familiarise themselves with current recommendations and how any serious breaches of the code might affect our appraisal of any outcomes (Ethical principles of psychologists and code of conduct, *American Psychologist,* 1992, 47, 1597–1611).

What potential threats of reactivity do you see with respect to (i) observer/scorer effects and subject expectancies, and (ii) observer/scorer bias?

In this section of the paper, we read that subjects "had something done" to them. This might be a special treatment, a new methodology, a data-collecting

test, an observation of their behaviour, a probing of their attitudes and habits, and so on. The researcher hopes that his or her subjects will "react" in some way to the instrument, method, or observation. However, this very reaction can often extend beyond the strict confines of the immediate objective and affect the very relationship between the observer/scorer and the observed. Such potential for "reactivity" threats will need to be considered by the reader, both in terms of possible observer/scorer effects and subject expectancies, and observer/scorer bias.

With respect to the first set of threats, research has shown that subjects may change their behaviour when they know they are being observed or assessed, and such behaviour may then no longer be typical of how they would normally behave. We might also imagine that, when the person doing the observing is also the person doing the evaluating, such changes may become even more apparent. Indeed, the reader might be even more cautious when he or she is told that observer, evaluator, researcher, and class teacher coincide in one and the same person! After all, our own experience might tell us that few class members would choose to react in a way that their teacher did not "approve of". Although these reactions are of interest as socio-psychological phenomena, they represent threats to the validity of data obtained through such observations, and the researcher would, at least, need to have been aware of them and, ideally, to have prepared for them in some way. The reader needs once again to be alert to the potential for this kind of threat to internal validity and weigh up the consequences accordingly. After all, by behaving in a way they think is suitable, subjects may unwittingly end up making a particular teaching method, instrument, or other variable look more effective than it really is.

The reader, in turn, needs to think about the way in which subjects might be reacting to the research situation. Subjects become aware that they are participating in an investigation, want to "do their best", and be a good contributor — all of which can translate into their behaving in the way they think the observer/scorer wants them to behave. Similarly, in a phenomenon known as the "**Hawthorne Effect**", subjects react in a way that is related to their pleasure at being included in a study rather than to any treatment involved. We might be particularly attentive to situations wherein intact classes are used, and one group is chosen (and is aware of the fact) to receive a new or special treatment or methodology while their companions in a similar (control) class continue with the "normal" input. We should also be conscious of the latent possibility for contamination as a result of such a situation. In the latter case, the danger would be that members of the control group with friends in the

treatment group may use the information acquired to change their own behaviour — and thereby the data collected from them. One obvious way of meeting this particular threat would be to use groups who are unlikely to intercommunicate, such as those in different educational institutions. As a direct consequence of the physical proximity of groups involved (i.e., within the same building), there is scope for communication of information about the study between groups of participants. Some of the potentially confounding consequences we should bear in mind in such contexts are that there might be a certain resentment on the part of the control group or rivalry set up between the groups (which may affect comparison data between them). Reactions of this type inevitably affect both the internal validity of what is being described and also the external validity of the same, since subjects would be behaving in a manner which is atypical of how they would normally react outside the experimental context. Likewise, such reactivity can also make a variable under discussion look more influential than it actually is.[15]

In our appraisal of the possible threats present, we might usefully focus on what we are told (or not told) about subjects' knowledge of the study itself and their roles within it. The argument would be that, by limiting subjects' knowledge about the aims of a study, more typical behaviour will subsequently be observed. Conversely, there is a fine line to be drawn between keeping subjects unaware of the objectives of observation and raising ethical objections by deliberately misinforming them about what is going on. There are, however, a number of other steps that a researcher might take that could increase our confidence in the way these threats were being met in the study. Firstly, he or she might use less obvious means of observation: in practice, this might mean the observer/scorer being outside the direct sight of the subjects or microphones and video cameras being made less obtrusive (but see below, "environmental conditions"). Secondly, subjects can be trained before the study commences to adapt to the observer/scorer's presence. The assumption here is that, given enough time, subjects will get used to this person and begin to behave in a more typical way in their presence. Finally, the researcher might think about using other indirect means of observation to supplement (rather than replace) direct observations and thereby help confirm such data: these can include homework, school reports, teacher's comments, individual learning diaries, and so on.

---

15. The reader is referred to the excellent discussion of contamination as a threat to internal validity in Cook, T., and Campbell, D. 1979. *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

At the other end of the reactivity relationship, we will need to assess the possibilities for observer/scorer bias. This comes about when this person — it might be the researcher or any of their fellow-observers — has certain expectations about the behaviour they are about to observe or record. This then leads to errors or bias in the correct identification and recording of this behaviour. Similarly, this observer/scorer may consciously or otherwise select the behaviour to be recorded. Logically, the observer or scorer who might therefore be said to be most at risk is the researcher. He or she is only too aware of the hypotheses or research questions at issue in the study and may have their observation biased by such knowledge. Moreover, other observers may also have become aware of the aim of the study in conversation or training with the researcher. We will, therefore, need to assess two possible sources of observer/scorer bias in particular in the kind of studies in our field: that resulting from the observation itself and that resulting from the recording of the observation.[16]

All perceptual processes involving observation, and its subsequent processing, will be subject to bias. It follows that observer/scorer bias cannot be eliminated, only moderated in some way. Some research contexts, however, are more vulnerable than others to this threat. The problem for the researcher and the critical reader to decide is how far data in the study in question are potentially or seriously threatened. Much will depend here on the instructions or training given the observer/scorer. Attention is normally selective in nature and is related to our present interests, experience, and expectations. Unless trained otherwise, observers are unlikely to make a conscious effort to distribute their attention widely and evenly. Readers might also pay attention to the amount of time said to have elapsed between observation and recording of data: the longer the delay after the event, the more likely it is that the recording will suffer in terms of precision and comprehensiveness. We might want to check on the instrument provided the observers to see how far it allows for on-the-spot recording.

A rather more subtle relationship may evolve between the observer/scorer and the instrument used to observe or score. Again, our attention as critical readers will need to be on the period of time during which the observer uses the instrument. Tiredness, boredom, or lack of attention are to be considered

16. For a more detailed discussion of how observer and observing bias may affect outcomes, the reader is referred to Rosenthal, R. 1976. *Experimenter effects in experimental research*. New York: Irvington.

normal consequences of such observation. As we read through questionnaire responses, listen to and categorise answers given in an interview, or evaluate compositions from large groups of subjects, our judgements inevitably become less focused, and the behaviour we are observing begins to standardise before our eyes. As a result, however, our judgements might become invalid: suddenly a specific response from a subject might cue a different reaction from the observer, an error previously considered serious is ignored or scored less punitively, or a particular behaviour is categorised differently than it was half-an-hour before. Such "observer drift" can be combated by introducing periodic inter-observer agreement checks, particularly since such drift is more likely as not to be an idiosyncratic process. We might be also looking out for the introduction of breaks in evaluation or the use of more than one observer/scorer to combat these kinds of threats.

Further guarantees might need to be built in to studies where observers/scorers are informed about the objectives of the study and are aware of the assignment of experimental and control groups. Again, judgements might be affected when an observer/scorer is conscious of the fact that he or she is being asked to assess someone who has "benefited" from a special treatment. We should be looking to see how far the researcher sought to offset this bias by, perhaps, limiting the amount of information provided them about the study. The researcher might try to use observers/scorers who are unconnected to the present study and do not know the subjects personally or academically. Furthermore, in a process often referred to as the "double-blind" technique, they could try to ensure that neither subjects nor scorers are aware of who is receiving the treatment. Unfortunately, the reality in the kind of typical small-scale study undertaken in our field is that such guarantees are difficult to fulfil. Once again, therefore, the reader will need to weigh up carefully the possible threat to data represented by this kind of bias.

(i) What details are provided about the environmental conditions of the study and could these have affected outcomes? (ii) What observations about external validity can be made in the light of these details?

A final set of factors affecting internal validity of procedures would have the reader concentrating on the environmental conditions of the study. In particular, careful thought needs to be paid to the physical arrangements for the study to see whether any natural conditions could possibly have impinged on outcomes. Again, it is doubtful that the researcher will describe such phenomena; the reader will need to consider the potential for such influence and any likely

consequences for the study. Such threats are of particular importance where studies are carried out in classrooms, and where conditions both for teaching and observation can be expected to vary greatly from one context to another. Although not, strictly speaking, vital data for replication purposes, information about the conditions in the "laboratory" help the reader picture the proceedings and think about whether any naturally-occurring phenomena might normally accompany such a setting. Much of this appraisal will need to be based on common sense and on our own experience of similar environments. Imagine, for example, that a comparison of listening comprehension abilities was set up between two groups after a period of treatment. Subjects were to hear a taped conversation between two individuals and subsequently answer questions on what they heard. Naturally enough, we would want to be assured that both groups were able to do the exercise in the same conditions. However, we would be especially interested in what those conditions were, since the validity of any data from the instrument itself (i.e., the tape recording and subsequent questions) would be particularly susceptible in this case to environmental phenomena such as outside noise, poor acoustic conditions as a result of the classroom design, or unfavourable seating arrangements (e.g., perhaps some subjects were too far away from the loudspeakers to hear adequately).

Similarly, we might want to consider for appraisal here other factors that may form part of the local set up for the research. How far do we feel the content and context of any passage read or heard might be familiar enough to the subjects to enable them easily to recognise what is going on? The reader might also want to know about any arrangements that impinge on the environment of subject response on the measure. These do not have to be only those which are naturally-occurring. For example, do the requirements of the task mean that subjects have to react "atypically" — perhaps having to stop and consider a number of options instead of replying immediately? Do they have to read or hear text in the L2 but react in the L1? Does the task require subjects to do two or more things at once that would normally require special training? Once again, our attention might be drawn to any factors that might reasonably affect or explain outcomes in the research other than the variables involved.

Indeed, these very classroom conditions can create a context which also threatens external validity in the study. Remember that externally valid data might allow the researcher to generalise beyond the immediate context of the study. In order to do this, of course, he or she needs to be satisfied that conditions within the study are similar, if not equal, to conditions that could be expected outside the study. The question then raised is "How real or artificial

are the experimental conditions in the study?". This is a particularly knotty problem in the field of research in second language learning. After all, how authentic a situation is practising a second or foreign language in the classroom? If we read that a researcher aims to describe subjects' L2 writing processes by having them think aloud as they write, we might ask how far data obtained in this way really succeed in probing the authentic experience of writing in another language. Either way, we might want the researcher to tell us how he or she prepared the best possible conditions and what conditions were actually like in the event. Secondly, we should return to the question of how time was used within the study. With regard now to external, rather than internal, validity, we need to consider how far the time periods set up within the study can be generalised to the outside context. The reader needs to judge here whether outcomes from a study in which time has been artificially restricted by the limits of the experimental period or the conditions of treatment can reasonably be externalised to a situation where such constraints may no longer apply. Second or foreign languages, for example, are not learnt in a day. Language learning is usually seen as a process in which we can only hope for lasting results after a considerable period of study. Although many teachers are happy to see that their students no longer commit a specific error after a lesson specifically directed at correcting this, few would argue that the real test comes much later when the students are working independently and speaking or writing in a less controlled situation. Once again, our interest should be drawn to research designs where improvement is claimed in an experimental group as a result of some particular treatment they received which the control group did not. The question to be asked is how far this perceived improvement could be a direct result of the short time period elapsed since the treatment was administered and, more importantly, how far this evidence of learning might then be sustained beyond the end of the study period.

Unfortunately, the list of potential threats to internal and external validity are many, and there are few simple solutions. Rather than our expecting to find control for every aspect of validity, it is sufficient that the research we are reading demonstrates an awareness on the part of the researcher that certain validity factors may have affected results and that subsequent interpretations of outcomes reflect this caution. In the end, the researcher will need to trade-off certain less-controllable factors for the control of others in the search for more valid — if less generalisable — results. For the reader, however, it is often harder to appraise the threats to external validity than to internal validity since the descriptions provided may cover internal conditions of the study to a far

greater extent than external ones. In the end, the onus will once again be on the reader to consider *all* such potential threats and weigh up their consequences for the research design and data analysis, as well as their influence on the confidence we can have in any reported outcomes.

2.3 Research design and data analysis

Identify the basic type of design employed here and draw the design box. What immediate observations can you make about this design and its consequences for the study?

As a result of what we have learned about the study in this and previous sections, we should now be better equipped to picture the research design employed therein. In many of the research papers read, where space is at a premium, we should not expect to be given great detail concerning the research design used. Indeed, advice about what or what should not be included in this section (and most others) is often dictated by journal editors. However, much of what we need to know should already have been made clear in previous sections. Being able to identify and then visualise a research design is a useful exercise for the critical reader for it will help us to clarify the appropriateness of the procedures carried out so far and put us in a better position to judge the suitability of any subsequent data analysis chosen.

What should firstly have become evident from the reading of the paper so far is the basic type of design. The identification of type will help the reader to ascertain the appropriateness of any proposed comparisons to be made between groups. Pre-experimental designs are simple and inexpensive to implement and exploratory in nature, but lack control groups to compare with the experimental group. They are often used in preliminary research to provide direction and focus for further research using experimental designs, or when circumstances exclude more controlled research design. In quasi-experimental designs, both control and experimental groups are used in the study, but subjects have not normally been randomly selected nor randomly assigned to these groups. In **pure experimental** designs there would have been prior random selection of subjects and random assignment to groups. In **ex post facto designs** the researcher studies the hypothesised link between two variables, but he or she is not interested in what went on before the study, and no special treatment is applied to the subjects.

Our previous understanding of the functions of the variables involved will help us now to make further important distinctions within the design identified. Thus, we will need to consider whether the study uses different subjects assigned to different (independent) groups (**between-group** designs) or uses the same subjects but in more than one treatment or taking samples at more than one time interval (**repeated measures design**). **Factorial designs** are used when the researcher wants to cross-compare, by crossing the levels of one treatment with all the levels of another. These designs try to separate out effects: they assess the main effect of each treatment and then the **interaction effects** of different treatments.

With this information, we can begin to draw a design box, which graphically represents the essential components of the design described so far. Although it may take time to do, by visualising the basic components of the study in this way, representing its exact particulars using the notation provided,[17] the critical reader will be able to spot possible weaknesses or anomalies in the design and consider workable improvements. The information so provided will also help us begin to see whether any proposed descriptive or inferential statistical procedures will or will not be suitable. First of all, write down the basic functional information we have been provided with about the independent and dependent variables, as in the following fictitious example:

You are a foreign language teaching assistant giving classes of L2 French conversation at a monolingual secondary school and a bi-lingual international school in Country F. These conversation classes are used as preparation for a section in the university entrance examination, which these subjects will be taking at the end of your six-month teaching period. You devise a questionnaire based on the proposed content of your course and its relationship to the examination. The aim is to collect subjects' opinions through questionnaire responses about how suitable or useful they think the course is going to be.

Independent variable: *Students of L2 French preparing for a university entrance examination.*

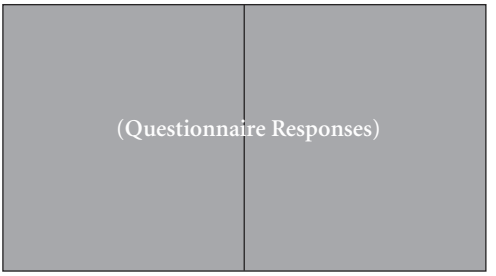
---

17. The drawing of design boxes is based on advice provided by Hatch, E. and Lazaraton, A. in *The Research Manual* (1991), New York: Newbury House Publishers. For the subsequent diagrammatic versions of research designs, I have used the classic notation suggested by Campbell, D. and Stanley, J. (1966), "Experimental and quasi-experimental designs for research on teaching" in Gage, N. (Ed.), *Handbook of research on teaching,* Chicago: Rand McNally (pp. 1–76).

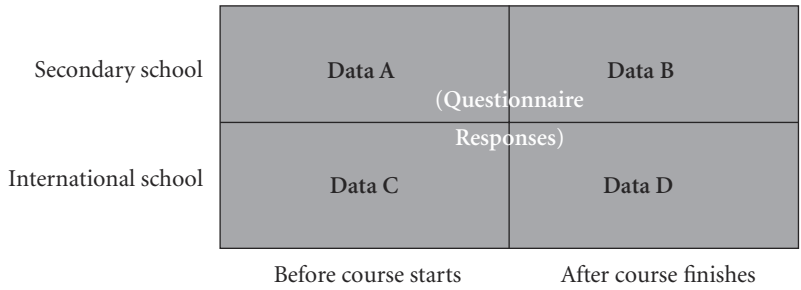
Level 1: *Secondary School*
Level 2: *International school*
Dependent variable: *Questionnaire responses.*

Research question: *What do these subjects think about the suitability and usefulness of the proposed course of study with regard to the university entrance examination?*

Now draw a box with the details of the dependent variable filling that box and the information about the independent variable at the bottom:



At this stage, we are going to be reading about two sets of responses which might, or might not, be compared (the research question does not clarify this). Furthermore, this is a between-group design, since two independent groups from two different schools are involved. We assume that members of the secondary school will not also be providing data in the cell wherein we find questionnaire data from the international school. Let us now imagine that we want to see if students continue with the same opinions at the end of our course. We administer the same questionnaire again to the same students and then compare what is found. Now the design changes somewhat: we have two samples taken at two different times and from two different schools.



This has now become a repeated-measures design: a sample of data (Data A) will be taken from the secondary school before the course starts and a second sample from the same subjects at the same school on the same measure, but at the end of the course (Data B). The same goes for Data C and D from the international school. Repeated-measures designs are very common in research in our field, presumably because of their immediate attraction in terms of procedures and selection. The subjects undergoing the treatments — or whatever happens in the experiment — have the same background, environment, age, gender, personality, and so on. However, what we also need to recall in our reading of the procedure is that the same individual has more than one involvement in the study: in cases where the same subjects receive more than one "treatment", there will be a theoretical effect of order on any measurements. Thus, whatever the subject experiences first will affect what happens next: this could present itself as a practice or "carry-over" effect from one "treatment" to the next (see p. 55).

Our appraisal can also show us the potential in a design for further study: the box helps us to see the theoretical possibility (i.e., depending on other features of the design, including selection and threats to internal and external validity) for a comparison between groups' questionnaire responses: in other words, we might go on to compare the secondary school responses before the study (Data A) with those of the international school (Data C) and so on. If such a design were feasible, the reader would need to be alert to the two kinds of statistical analyses being carried out (i.e., between-groups and repeated-measures) since both require slightly different operations which, if carried out wrongly, can lead to what is known as a Type 1 or Type 2 errors.

Visualising designs in this way is also useful in appraisal since it helps us to spot inherent anomalies which could end up confounding that design and, therefore, invalidating any outcomes. Let's go back to our hard-working foreign language assistant:

As you are preparing both sets of subjects for the same examination, you decide to try out, and compare, a new teaching approach to the examination. You decide there is no possibility for contamination or communication between the groups because of the physical distance between the two and you think the two groups would be similar in terms of age and L2 proficiency levels. You toss a coin to decide which of the two groups will receive the new approach (consisting of structured debates in pairs, which are recorded and discussed in tutorials between the pairs and the teacher). As a result, the international school subjects are assigned to the new approach, while the secondary school subjects get the recommended teaching approach (consisting of class discussions and individual projects delivered orally before the whole class). You give both classes the same test at the end of the six-month period to see if there is any difference in results.

Now let us draw the new box with the following information about the variables and the research question:

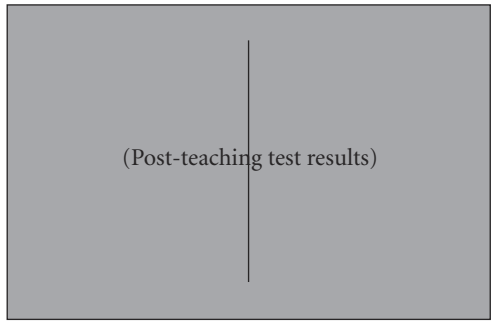
Independent variable 1: *Students of L2 French preparing for a university entrance examination.*
Level 1: *Secondary school*
Level 2: *International school*

Independent variable 2: *Teaching method*
Level 1: *Traditional*
Level 2: *Experimental*

Dependent variable: *Post-teaching test results.*

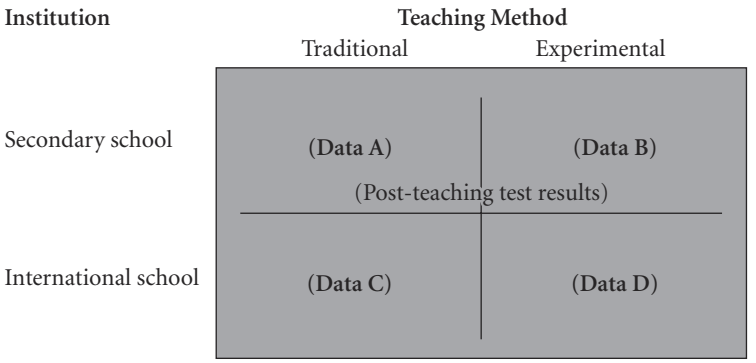
Research question: *Is there any difference in test results between subjects who have received the new teaching method and those who received the traditional method of teaching?*

The text seems to indicate a "simple" comparison between groups, which will establish whether one treatment brings about a difference in outcomes. Setting aside the many possible threats to internal validity here for a moment, let us draw a box with the details of dependent variable filling that box and the information we have been given here about the independent variable and its levels:



Once the assistant collected the data, he or she could then compare the treatments and prepare to achieve instant fame if their new approach is shown to "work". However, the design box drawing reveals a much more serious problem. There are two independent variables present. Any differences between the groups might well be due to the teaching approach received; however, they might also be due to the educational institution variable (i.e., secondary school teaching context versus international school teaching context). Hence, the design box drawing succeeds in bringing to light a confounded research design. Confounding is possible when a researcher allows two potentially effective variables to co-vary simultaneously. While some interesting insights might be revealed as a result of *describing* what happened in each institution here, the assistant will not be able to attribute differences in test results at the end of the teaching period to either the teaching approach or, for that matter, the educational institution involved.

Although we may read of more complicated (or complicated-sounding) designs, it is always a useful test of appraisal to try to draw the basic design. Apart from revealing possible confounding, the exercise might also expose other potential relationships that the researcher may not have seen, or may not have considered interesting for their purposes, but which we find noteworthy for our own future research. Imagine, for example, that our reading of a similar paper produces the following box diagram, which represents a factorial research design:



We are presented with two independent variables. The researcher, let us say, still claims to be interested in seeing whether subjects taught using the experimental method do better than those who receive the traditional method. We again read that the dependent variable is the post-teaching test results. The first independent variable is "Teaching method", and there are two methods involved here. The researcher also wonders, however, whether subjects from the two institutions might perform in a different way on the test. There is no longer any confounding inherent in this box design (there are four classes involved), so we can proceed to discuss its implications. This "2×2" matrix or factorial design sees the comparison of two methods and two schools. Consequently, four basic comparisons would be open to the researcher, who may or may not pursue all of them. He or she could compare how all the subjects in the secondary school with all the subjects in the international school (Data A and B vs. Data C and D). We then read on further in the paper and discover that the researcher also wants to compare all the subjects receiving the traditional method with all the subjects receiving the experimental method (Data A and C vs. Data B and D). The researcher decides to stop there, but other combinations within this design could be studied and, we might suggest in our appraisal, *should* have been studied at the same time for the light they might throw on the interaction effects of the different variables. It is possible, for example, that a specific teaching method may help subjects more in one institution than in another; in other words, one or other group may profit more from one approach than another (Data A vs. B vs. C vs. D).

The advantage of these designs for the reader, and of drawing their visual representation, is that a number of interactions can be set up and investigated or suggested for future research. Here, for example, we might read in the hypothesis that the researcher suggests subjects in the secondary and international

schools will normally perform in the same way and that the only differences come from the experimental teaching approach itself. Nevertheless, the reader would be looking for data from the researcher which confirmed that no other "undesirable" interactions were taking place that weakened the above hypothesis. For example, subjects receiving the traditional method in the international school (Data C) might have done better than those receiving that method in the secondary school (Data A). If these data were not forthcoming, our appraisal of the paper might include the need for replication of the study with a different focus for analysis.

Attempt visually to classify the data-collection procedure, and comment on the perceived consequences of this for any eventual findings. Where necessary, suggest how this might have been improved, and why.

Once the basic design of the study has been established, it is then a useful exercise to try visually to classify the proposed procedure into one of the recognised design classifications for this kind of research. A number of such classifications exist for experimental and quasi-experimental research, and for each classification there are potential threats to reliability and validity to be considered. As with so much in our appraisal of others' work, we must assume it will be the responsibility of the researcher to justify the classification he or she has opted for and acceptably address any such weaknesses. In practice, however, for reasons of editorial policy, space, or perceived priorities, he or she might not be explicit in this regard, and it is again left to the reader to decide on the adequacy and appropriateness of the classification in the present context. The opportunity should not be wasted, since drawing and classifying the research design will again help us to evaluate a key element of the study and will allow us, firstly, to comment on the degree to which this contributes to or detracts from the confidence we can place in any eventual outcomes and, secondly, to suggest possible future improvements to the design.

Before we embark upon classifying the design, the reader needs to check on a number of basic elements of the design so far. Firstly, a check should be made to see whether a pre-test was administered. Then, we should be interested in whether random assignment was used to assign subjects to groups, whether intact classes were used, or, indeed, whether one class was used and received a period of treatment followed by lack of treatment (see below, "time-series" designs). At this point we will have the information required to confirm the design classification and whether this is going to be basically a pre-experimental, quasi-experimental, pure (true) experimental or *ex post facto* design.

Finally, it will be useful to check on whether a factorial design is being used by establishing the number and nature of any independent variables. Using the accepted notation, we should now be in a position to draw a representation of the study as it has been described.

It is beyond the scope of this book to illustrate and discuss all the possible design classifications and their respective modifications associated with each of the above design types, and the reader is referred to the specific books and manuals in the "Further Reading" list. However, in order to make an adequate appraisal, it is useful to understand the basic principles of classification, learn how to apply the notation to the more widely-used design classifications in our field, and examine how appropriately the studies we read correspond to such designs. To this end, what follows is intended both as a summary of the most salient points in some of the principal designs, and also an aid to addressing the appropriateness of the steps the researcher has chosen to take in setting up the study.

The recommended notation is as follows:

X means that a group was given or exposed to some experimental event, the effects of which will be measured.

O represents an observation or formal test measurement. A number in brackets afterwards indicates the number of observations made.

X and **O** in a particular row apply to the same subjects, while vertically presented means that they were applied concurrently.

The order of events is described through the left-to-right direction of the arrowed line.

R means that subjects were previously randomly assigned to the groups.

Pre-experimental designs

We have already highlighted the fact that much of the research in our field is shaped by the context in which we find the subjects we must use. So often only intact classes are available for study, and the researcher must accept and allow for the inevitable restrictions — many of which have been described in the course of this book — these participating classes place on the interpretation of any data obtained. This does not mean that such research designs (and the findings that emerge from them) are somehow inferior to true experimental

research.[18] Rather, little can be read into any outcomes. Thus, they will demand considerable replication, but through that replication the field might begin to discern tendencies. By systematically analysing such tendencies across replication studies, a picture will eventually emerge about what is really happening in a specific link between variables, and quasi- or pure experimental designs can then be set up to investigate these further. However, the reader and his or her projected research can still benefit from the immediate exercise of identifying the type of restrictions such designs place on data, and suggesting possible improvements with a view to improving any subsequent replications. In what follows, a number of suggestions are made to help us go about doing this.

The most basic design classification is one in which subjects are given a test on something that they have seen or received earlier: **X**→**O**. The appraisal of such a design would need to concentrate on what it does *not* tell us, rather than on what it does. While the researcher might adequately describe the subjects, the research context, and the variables involved, there is no information about the control of any possible extraneous variables. Scientific evidence for links between treatment and an observation is based on processing comparisons, not on isolated instances. We know nothing about the group's characteristics before the treatment was given; as a result, it will be almost impossible to infer any kind of effect. In other words, the researcher will not be able to conclude that **X** *caused* **O**. On the positive side, we would agree with the researcher providing a description of the data obtained; however, we would also be looking for a firm acknowledgement that there are serious enough threats to validity and reliability to recommend extreme caution in interpreting any results. The two obvious suggestions in our appraisal for improvement would be to include some element of pre-testing and/or a control group (i.e., a group that does not receive the "treatment"):

O(1) X O(2)
→

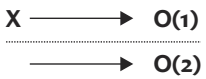
By introducing a pre-test we do gain some information about the sample. The researcher would then be able to assure the reader that the students did not already know the material tested on the post-test, for example. A control group would not be needed as the group would classify as its own control. There remain,

18. Other scholars can be more critical of these designs, highlighting the inadequate nature of their validity (e.g. Tuckman, B. 1994; Robson, C. 1993. *Real World Research.* Oxford: Blackwell).

however, a number of disadvantages to the design which would need to be taken into account in any subsequent appraisal. It fails to control for threats from history or maturation factors. Could not other events have taken place between **O(1)** and **O(2)**, apart from **X**, to produce the effect or relationship observed?[19] We would be thinking of events that affected the whole group, of course. Indeed, the longer the time lapse between **O(1)** and **O(2)**, the more plausible such "history" factors might become. We might be interested to read about what the "normal" class procedures and content were during the study for these classes. It is, after all, possible that they would have improved anyway given this normality and the time period involved. Furthermore, the illustration also warns us about the possible effects of timing of events: we would need to read the procedures carefully to see if the time elapsing between the pre-test and the post-test was particularly short, or if the pre-test content and questions inadvertently "warned" subjects about the subsequent treatment and post-test. We might also consider any evidence that these in-between periods of time or parallel events led to those observing or evaluating somehow changing (i.e., the people themselves or their way of observing/evaluating) by the final observation/measurement.

Quasi-experimental designs

The great disadvantage in the above designs, therefore, is that they do not manage satisfactorily to eliminate alternative explanations of outcomes. Fortunately, in the majority of cases, we have seen that a little thought before designing the research can produce much more valid designs and outcomes in which the reader can have more confidence.



Quasi-experimental designs are often used in our field since — while we are able to introduce certain elements of experimental design into many of our studies — we often lack full control over various aspects of the procedures. In

19. Robson (1993:101, op cit.) points out that — under certain, highly controlled, circumstances — such a design can become more interpretable. For example, the researcher might have been able to isolate the group from outside influences or to have demonstrated the absence of any pre-treatment tendencies.

the above classification, a control group has been introduced, which consists of another class. However, randomisation has not been used to assign classes to experimental or control status. This design is one that we may encounter quite frequently in research carried out in educational institutions, and may be thought to straddle a "pre-experimental" and a "quasi-experimental" design. There is a definite improvement on the above designs, since a second, "control" class has been added. Its principal attraction in such a context is the fact that it makes use of available groups and can establish a comparison group without "disruption" of the school system and/or re-assignment of subjects to other classes than those in which they were originally placed. According to the design, a treatment is given one group, and then its results are compared to a second group who have not received the treatment. It would appear that threats from history and maturation factors could be controlled somewhat by the use of the control group. If a coincidental event within the school procedure affected one group, it would probably affect the other as well.

Faced with this kind of design, the reader would again need to have their attention drawn to the serious limitations in the interpretation of any results. We know nothing about the similarity of the two groups *before* the treatment started; any improvement noticed in the experimental group might also be down to a number of co-existing extraneous variables as well as to the treatment itself (e.g., dissimilar L2 proficiency level, different time of classes during the day, different teachers involved, or subject gender). Recommendations for future improvements might include the random assignment to groups and/or the introduction of pre-testing since, as it stands, it is impossible to discover if the two classes really were comparable to start with. Some more validity might be also gained by recommending that future replications begin by matching subjects between the two groups for key characteristics such as gender or assessed language proficiency scores. When randomised design is not feasible, matching subjects will go some way to permitting increased comparability between the two groups (although the reader will still need to consider the possibilities for "regression to the mean" (see p.41).

Faced by the difficulties in obtaining random selection and assignment of subjects in many educational establishments, together with the evident drawbacks of pre-experimental designs, many researchers opt for designs that follow a single group over a period of time (often referred to as a "time-series" design). Again, intact classes would be used, but this time the class continually serves as it own control:

O(1)-O(2)-O(3)-O(4)-X-O(5)-O(6)-O(7)-O(8)

The advantage for the researcher here over the second pre-experimental classification above is that, now, several pre- and post-tests are administered prior to and subsequent to the treatment itself. Furthermore, only one class is involved, and there is no necessity to find matching subjects in another control group — all of which makes for considerable convenience in many research contexts. Our confidence in any outcomes should also be reinforced by the fact that the researcher meets threats to history by taking a number of samples throughout the study. Once a particular pattern of behaviour has emerged before the treatment, the latter can be applied with confidence and the subsequent samples examined for any change. The design gives the researcher the chance to detect change at various points after the treatment, such as immediately following the treatment (**O(5)**) or at successive samples. In general, any possible threats from unconnected events in everyday class (or home) procedures should be seen across all the observations and might then be adequately accounted for across the whole study.

Faced with data from such a study, the critical reader would still need to look carefully at a number of aspects. Firstly, careful consideration is needed of the number of observations or samples obtained so that we consider the evidence of growth trends or change after the treatment. With more points of data collection, the researcher would be in a stronger position to assess the nature of the trend. The reader, however, should be alert to whether the trend post-treatment contains any apparent anomalies. Perhaps, instead of increasing immediately after the treatment, no increase is noted until observation 7. This would indicate a trend we would want to see addressed by the researcher. The same would be true in the case of a non-linear trend that increased immediately after the treatment, but then decreased suddenly, or subsequently revealed much smaller increases. Furthermore, careful attention needs to be paid to any reported gradual, rather than abrupt, changes after treatment, since the possibility exists that these are merely normal fluctuations over time (see Cook, T. and Campbell, D. 1979. *Quasi-experimentation: design and analysis issues for field settings*, Chicago: Rand McNally).

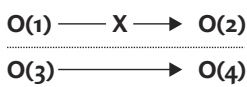
A second important aspect to consider here would be the kind of data used to obtain the pre-treatment observations. For example, a researcher may look to available school assessment results. In this case, such material will need to have received adequate scrutiny on the part of the researcher to establish its reliability, validity, and general appropriateness to the study at hand. After all,

we would hardly want to accept assessment of listening comprehension proficiency as adequate measurement of overall L2 proficiency in such a time-series design. Indeed, this is true for any freely-available "pre-experiment" data offered the researcher and then used in the study. Generally, such data will have been obtained for totally different purposes, which may mean that it is systematically biased or has been collected according to different criteria at different times and by different people.

The time period itself might usefully form a subject of discussion: we might want to check on the length of time the group was studied and think about any parallel activities in which the subjects were thought to participate and which might have affected results. Longitudinal studies such as these typically extend over time periods that involve seasonal change, variations in timetables or work-loads, all of which can potentially be confused with change (or lack of change) resulting from a particular treatment. Further recommendations for improvement might include the setting up of a comparable control group who are observed at the same time and the same way, but who do not receive the treatment. The main advantage here is that this group would then further test for any "history" threats. Secondly, we might suggest that such threats could be countered by using the group as its own control, and taking observations on a second dependent variable which should not be affected by the treatment.

This same design would also prove an adequate basis for any single-subject experimental design, although it has the potential problem that the researcher cannot then compare any trends across other subjects.

Another commonly-encountered classification in our field is one which builds up on, and strengthens, a previous pre-experimental procedure. Once again, a control group may have been found by using another class from the same section. While subjects have not been randomly selected for the course, nor randomly assigned to their respective sections, the researcher still has the chance to randomly assign the experimental and control groups on the toss of a coin:



The improvements brought in to this design should enable us to have more confidence in any eventual findings. There is a control group, pre- and post-test measures for both groups. In the case of non-randomisation, pre-tests are essential, unlike in the pure experimental version (see below), to assess the group similarity on the dependent measure. The reader, however, will still want to examine the details of such pre-testing carefully. As we suggested above, by

adding a comparison group that is truly comparable and has similar experiences to the experimental group (apart from the treatment), the researcher is theoretically able to answer most possible threats to history, maturation, testing effect, and instrumentation. We might assume that both groups change naturally at the same pace, experience the same effect of testing, or are exposed to similar external events. Assuming these events *are* experienced in the same way by both groups, they will not intervene on any post-test measurements.

Despite these increased safeguards, however, this will not obviate the reader's need to stop and consider any pattern of results reported from such classifications. Let us take the example reviewed above. Imagine that we read of a six-month study in which the researcher has, indeed, set up an experimental and comparison group (intact classes), pre-tested both in the way suggested, provided the treatment to just one group, and finally post-tested the two groups. We read that pre-testing revealed the experimental group to be slightly higher on an important measure than the control group. On the post-test, there is an increase "as expected" in the experimental group; however, the same test reveals a much smaller increase in the comparison group. What might we consider here? It appears that the control group might be changing; a question we would want to see answered is whether the experimental group with the pre-test advantage is changing (maturing) at a faster rate regardless of the treatment received.

The cautious reader will still need to be alert to the characteristics of each component in the design classification of a study, so that he or she can judge whether any intentional or unintentional modifications to this design place undesirable threats on the eventual outcomes. For example, while general history bias threats might be countered here, we would need to be particularly attentive to designs wherein the experimental and control groups go through the same experiences but in *different* centres: problems of local history become more acute as the number of controllable group locations increases. Also, an adverse interaction might take place when, as a result of the treatment, subjects in one group grow considerably more experienced at a faster rate than the other group. When the experimental group consists of subjects who are brighter or more competent than those in the control group, differences in maturation may confound data. Similarly, the reader might wish to know whether the experimental group were self-selected (i.e., volunteers for the experiment). The motivational characteristics of volunteers is not typical of all subjects. In other words, comparing a group of volunteers to a group of non-volunteers does not control for internal validity because of the different inherent nature of each group.

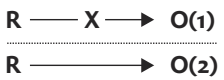
With the central importance attached to pre-testing in this classification, the reader may think it useful to know exactly what this consisted of, and who did the examining. Firstly, we need to recognise that, even when pre-test measurements profess the similarity between the two groups, we cannot assume that the two groups are necessarily equivalent. For example, by looking at the details of the pre-test, the reader should be able to confirm whether or not subjects have been tested on one or two measures of L2 proficiency only (for example, their writing ability and/or their ability to correct spoken or written mistakes). The fact that these subjects are similar on one or two such measures does not mean that they are equally similar on other measurements that may be just as relevant to the study at hand (for example, their ability to plan their writing or to attend to spoken production in order to spot mistakes). Secondly, we might again want to look for information about whether the pre-tests (and post-tests) of both groups were evaluated by the researcher and/or people who were aware of the nature of the two groups. Presumably, such awareness or "training" on behalf of those who do the measuring could affect the kind of measurements made.

Pure/True experimental designs

It has been my contention in this book that pure experimental studies are rare in our field because of the nature and context of much of our work. Such designs provide completely acceptable controls for all sources of internal validity. External validity — generalisability to the "real" world — is much easier to achieve outside the "laboratory" in a setting that corresponds to that "real" life. Such studies will use control groups, they will measure and, if necessary, control for differences before the treatment. Random selection of subjects will be possible and random assignment to groups carried out. Which group receives the experimental treatment and which becomes the control will also be based on a random decision. Given the comparative rarity of such classifications in our field, we will only look briefly at two typical designs here, both of which develop logically from their quasi-experimental and pre-experimental equivalents.[20]

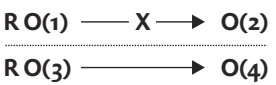
---

20. A large number of research methodology books deal with such designs in more detail. With specific regard to their appraisal as research designs, particularly useful accounts can be found in Cook, T. and Campbell, D. 1979. *Quasi-experimentation: design and analysis issues for field settings*, Chicago: Rand McNally; Campbell, D. and Stanley, J. 1966. "Experimental



This design improves considerably on the first quasi-experimental design described above. Grouping is done on a random basis, which basically controls for selection bias. Statistical analyses are used to determine the probability that the observed link between variables occurred by chance alone. If that probability is sufficiently low, the conclusion is made that it was the treatment, rather than pure chance, that caused the difference.

An obvious improvement will be the addition of a pre-test:



The pre-test gives the researcher the interesting option of matching individuals based on their pre-test scores or other criteria and then comparing the pair's performance after the treatment. Also, he or she could compare subjects' gain scores, rather than the final test scores, by simply subtracting the pre-test from the final test scores. This provides a useful means of taking into account individual differences and can therefore offer a more precise indication of treatment effects. Finally, if as a result of the pre-test the two groups prove not to be equivalent, it is possible to use a statistical adjustment to the pre-test measurements (such as ANCOVA) that may give a fairer picture of the treatment effect. Such procedures will all serve to strengthen the reader's confidence in the study by adequately meeting threats to internal and external validity. The reader will again need to consider details of the pre-testing very carefully. According to what it contains and when it was administered, a pre-test might sensitise the subjects in a way that their eventual post-test scores are affected.

Ex post facto designs

These designs are included in this brief tour of inspection because they are also very commonly used in our field. It is often too difficult to meet the many

---

and quasi-experimental designs for research on teaching" in Gage, N. (Ed.), *Handbook of research on teaching* (pp. 1–76), Chicago: Rand McNally; and Keren, C., and Lewis, C. (Eds.) 1993. *A handbook for data analysis in the behavioural sciences.* Hillsdale, NJ: Lawrence Erlbaum.

threats to internal and external validity and, therefore, it becomes highly inadvisable to make claims of cause and effect as a result of our research. *Ex post facto* designs enable the researcher to study the hypothesised link between two variables, but he or she is not interested in what went on before the study, and no special treatment is applied to the subjects. The popularity of this kind of design in our field is easy to understand. Our knowledge of the process of second-language acquisition has increased greatly over the last three decades. However, there is still much more for us to learn and such discovery will be based primarily on finding out what is actually happening in the process, rather than "intruding" on that process. When we have built up enough information concerning what is actually going on in second-language acquisition, we will be in a better position to begin to see how to improve things by intervention and experiment in that process. Hence, there is a continuing need for studies from different learning contexts that tell us what is currently going on in the L2 learning process. This will provide a useful opportunity for the researcher to describe some data and study how these change (or do not) across different contexts, subjects, and tasks.

The important thing for the reader to remember when presented with such a design is that the researcher is not interested in seeing the effect of a treatment as such, but rather in studying the hypothesised effect of an independent variable on another after that effect has "occurred" (such as whether gender has any effect on motivation in second-language learning).

Although this kind of research may not always be visualised differently from the other designs in this chapter, the main distinction is that any "treatment" is already present by nature rather than created or manipulated by the researcher.

O(1) **O(2)**

Here the researcher will obtain two (or more) sets of data from subjects with the aim of describing the link between those data. The research is not attempting to show that performance has improved as a result of some instruction or other, nor is cause and effect being studied, no group assignment is organised or needed, and no variables are being manipulated to bring about a change.

In appraising results produced from such a study, the reader should be alert to the interpretation of any observed relationships between both observations or measurements. There is a temptation to interpret this design as quasi-experimental and then suggest that the variable measured by **O(1)** has in some way brought about the changes observed in the variable that **O(2)** has measured. However, we should remember that a number of possible causes of this

outcome can be posited: firstly, **O(1)** may indeed have brought about what happened in **O(2)**; secondly, it is also possible that the variable **O(2)** is measuring had caused **O(1)**; finally, some other, unspecified and unidentified, variable has caused both the outcomes at **O(1)** and **O(2)**. We should remember that no manipulation of variables has gone on here and no treatment is administered, which means that none of the above explanations is confirmed. A suggestion — nothing more — of a relationship exists. The logical call from the researcher would be for further replication of their study to confirm the suggestion, and/or perhaps an experimental or quasi-experimental study which applies a suitable treatment.

Another type of design is applied when description is needed and the researcher is able to contrast one group of subjects who are said — or observed — to possess the characteristics under discussion with another group who have the "opposite" characteristics. This criterion group design can be used to establish how this behaviour came about. For example, how do students who are said to be "good language learners" get to become that way? What experiences have they had, and what training did they receive? Again, the reader will need to bear in mind that causal accounts of this experience or training will not be acceptable as an explanation of what is observed. The best that can be hoped for are indications of links, which will then need to be subjected to more rigorous study. Similarly, the same design classification might be used to describe the way members of different criterion groups (e.g., "poor L2 listeners" and "good L2 listeners") behave in a similar language-learning situation (e.g., do the latter criterion group make more use of paralinguistic clues when they have to understand and participate in an L2 conversation in a busy street?). Again, the design will not permit conclusions about cause and effect because factors other than listening skills as such might be bringing about any differences observed. However, a valuable platform will have been created for future study of the association between variables, in this case listening ability and paralinguistic clues.

Factorial designs

These designs have much in common with true experimental designs. Essentially, they can make use of the basic elements found therein (such as randomising selection and group assignment, pre- and post-tests) but with the modification that the effects of a number of independent variables (realised as moderator

variables) are being measured. Thus, within the factorial design, more than one variable can be manipulated and studied.

As we noticed in the basic design box above (p. 69), the factorial design can involve all possible combinations of the levels of the different independent variables. Thus, a fairly simple factorial design with three independent variables, one having two levels, one having three levels, and the other five would have a theoretical set of thirty (i.e., $2 \times 3 \times 5$) combinations! Conceptually, there is an unlimited number of complex designs, because any number of independent variables can be studied and each one can have any number of levels. As we shall see in subsequent sections when the actual analysis is carried out, this very "multi-level" characteristic can easily get out of control, with data being reported on any and every combination available, be it of interest to the research hypothesis or not.

The most straightforward design involves (as we saw in the design box) two independent variables, each of which is manipulated at two levels. As a result of these groupings, a number of combinations emerge for potential study. It will be possible to assess the separate effect of each independent variable (known as the "main effect") as well as whether the effect of one variable differs from one level to the other of the second variable (known as "interaction effects"). In this way, a study can report on the fact that one of these variables actually moderates another to achieve a certain effect. A total of four groups are involved in this example. The factorial design helps the researcher to "cover more ground", as it were, in one study: there is an assessment made of the effects of two variables, and also of whether they interact.

Analysis procedures for appraising the **statistical significance** of main effects will be described below and in the "Results" section, but suffice to note here that the interpretation of the main effects is totally dependent on whether an interaction is or is not present. In general, main effects will need to be interpreted with great caution whenever an interaction is reported in the study. When no such interaction is evident, the main effects of each independent variable can be interpreted as though they had been manipulated in two separate studies, each involving only one independent variable.

> **What procedures are identified for data analysis, and do these deal adequately with the original objectives of the study? In the absence of information about procedures, suggest how this might be done.**

In the data analysis section of the paper, the reader would be looking to see what was done to the data once they had been collected in the way described in the

previous section. Some sources recommend that this section be embedded in the "Results" (see, for example, the *APA Publication Manual*). However, I will separate them for appraisal here as there are a number of specific features of each which — for these purposes — are best considered apart.

Now that we know what the research question or hypothesis is, what variables are involved, how these are to be measured, how they are hypothesised to be linked, how these are to be operationalised in data, what groups and materials are involved, and how all these are to be used in the research design, we now have the tools to consider the appropriateness of the kind of data analysis proposed in the paper. Careful attention will now need to be paid to how, and why, the researcher opts for this or another analysis as well as to the sequence used in the process of analysis. Where descriptive or inferential statistical procedures are to be used, a major concern of the reader will be to consider, firstly, the appropriateness of such a procedure in the present research context and, secondly, whether the necessary "rules" or assumptions associated with each procedure have been adequately met both to permit such an analysis to be made, and also to obtain results in which we can have confidence. Once again, the specific publishing medium will often have determined the kind, and amount, of detail to be included about analyses. For example, the journal *Tesol Quarterly* concentrates on consumers of the research and recommends (in its "Guidelines for submission") authors include enough information "to allow readers to evaluate the claims made". The *APA Publication Manual* reminds the author directly that analyses should be reported in enough detail to justify any conclusions made later.

If the statistical tests used are common ones (such as *t*-tests, ***chi-squared***,[21] analysis of variance, or correlation), the step-by-step details are often omitted in the text. Only the name of the procedure will be given and it will up to the reader to be in a position to consider the correct application in each case. Since our immediate objective is the appraisal of this section, and this requires our focussing on several of the steps involved in these tests and the principle assumptions behind them, the following summary is intended as a review of the key elements to which the reader should pay particular attention in their evaluation of the section. In this and subsequent sections of the book, it is assumed that the reader has some prior acquaintance with the basic concepts and language of descriptive or inferential statistical analysis. For a full description

21. Also known as "chi-square"

of how to perform these step-by-step statistical analyses, the reader is referred to the manuals in the "Further reading" list.

Try firstly to determine what tests or procedures have been specifically identified by the researcher to analyse the data. The procedure(s) selected could then be immediately referred back to the hypothesis or research question to confirm how far the data bearing on all relationships or comparisons posed therein are likely to be subjected to these analyses. Since the data analysis and results sections are often combined for questions of space, it may well be that the reader will need to indulge in the useful exercise of predicting and discussing, rather than identifying and analysing, the most appropriate procedures, based on what he or she has read in the paper so far. These deliberations would then need to be confirmed in the subsequent "Results" section.

Provide a step-by-step description of the elements involved in the data analysis so far, and decide on the appropriateness of any proposed analysis procedures in the light of this.

As so often before in our reading, the key word in this section is "confidence". Statistical analysis is used to give the researcher and the reader confidence in the claims being made for the data. However, the reader's confidence in the outcomes reported will again need to rest on more than just a set of numbers in the "Results" section of the paper. Frankly, any statistical test on data can be "made to" turn out significant; the computers that calculate this for us do not know, and do not care, how the data got there in the first place! The interested reader appraising this section of the research paper would want to delve deeper into what has happened prior to the feeding in of data into the machine and use what he or she has read to establish or confirm the ideal descriptive or inferential statistical procedure to be used in the circumstances.

Descriptive statistics do just, and only, that. They describe data in a way that allows the researcher to inform us about how often something occurred in the data, what typical values or elements were found in the outcomes, or how such values were dispersed throughout the data obtained. Typical statistics which the reader should be looking out for in such cases are some measure of **frequency**, **central tendency** (such as the **mean**, **mode**, or **median**), and variability (typically the variance or standard deviation). All three measures can provide important insights into data and help us understand them better. In most of the research described in the previous section, however, the researcher will want to do more than just describe. Descriptive statistics will tell us finite information about our particular sample of subjects; they do not, however, help

us to think beyond this sample to the larger population. As in the research we have appraised so far, most researchers undertaking experimental or quasi-experimental studies will want to know whether the data described can be used to support hypotheses made or help to provide responses to research questions. According to the *APA Publication Manual*, the eventual data presentation (i.e., in the "Results" section) should be reader-friendly in the sense that sufficient information is provided therein to permit the reader to confirm the appropriateness of the analyses and any information about differences in scores or frequencies or other measurements. But both reader *and* researcher will probably want to know whether these differences are normal or large enough to make certain inferences about them and their causes. Statistical tests can be used both to describe and to infer. When a researcher uses them to describe, such tests help him or her to have confidence in the description made of the data. When he or she uses these tests to generalise (inferential statistics) the idea is to make inferences from our data to other subjects and other learning contexts.

To begin to decide about the appropriateness of any proposed data analysis (or to predict what might be the most suitable analysis), it is useful to go through the same step-by-step assessment of the situation and objectives as the researcher should have done to arrive at the decision. There is a variety of choice open to the researcher for the analysis of data, but this will have been determined by the characteristics and measurement of the variables, the type of research problem, the research design, and the nature of the data obtained. We assume the researcher would have been seeking to analyse his or her data in a way which throws light on the research question or hypothesis. In the kind of research design that concerns us in this book, data will mostly be in original, or converted, numerical form, and the aim would be to submit this to some kind of descriptive or inferential statistical procedure.

The first data we will need to consider should already have been provided. Firstly, establish the number of independent and dependent variables and the number of levels of each variable. Secondly, confirm what kind of comparisons or relationships are sought: look back to the procedures section and check on whether the researcher intends to compare the one group of subjects with another group on one task or different tasks (between-groups) or with themselves on one or more measures (repeated-measures). Thirdly, we will need to establish the way the variables are to be measured. In other words, look back to the "materials" and/or "procedures" section and see if these are to be measured as frequencies (nominal), ordinal (ranked), or interval (score) data. At this

point, we can consult the flow chart[22] in the Appendix (p. 248) to establish the preliminary options available for analysis. This chart describes some of the most common descriptive or inferential statistical procedures.[23]

For example, let us imagine we are reading a study in which one group of beginner-level L2 students were trying out a new reading-comprehension programme, and another group were acting as controls. Data to assess reading comprehension proficiency will be obtained on a specially-designed in-house test. While the researcher is aware that she cannot use inferential statistics to generalise any results — amongst other reasons, because intact classes are being used — she would like to see whether the new material has any significant effect on test scores within her classes. Thus, we determine

Variables: Dependent = *Reading comprehension proficiency;*
Measurement = *scores on in-house test (interval)*
Independent = *Groups (two levels);*
Measurement = *nominal (experimental and control).*

Comparison or relationship to be tested = *Comparison between independent groups (between-groups).*

Turning to the flow chart, it is clear the researcher intends to "*discover the effect of an independent variable on a dependent variable*" here. She will "*measure the dependent variable*" from the in-house test (i.e., "scored" data). The chart now gives us three options, according to whether the design aims to use different, the same, or **mixed groups**. Here two different intact classes are involved so we choose "*between*" in the chart. At this point we must finally decide on the "*levels*" of variable involved. This study has two groups of one independent variable involved. The chart suggests "*t test*". Although the reader can use the chart on a reverse route (i.e., checking on the step-by-step logic behind a researcher's decision to opt for a stated procedure), more care is required here since a number of other assumptions (see below) may, and should, have gone into that researcher's decision.

---

22. The flow chart is reprinted from Hatch, E., and Lazaraton, A. 1991, *The Research Manual.* New York: Newbury House Publishers

23. Readers are recommended to consult the tables in Brown, J.D. 1992. Statistics as a foreign language: Part 2. *Tesol Quarterly*, 26, 4, pp. 636–637 for considerations about more advanced statistical procedures.

Once we are aware of the basic test recommended, the logical first step would be to compare this outcome with the decision made by the researcher, assuming that has been given in the text so far. If our selections do not coincide, of course, the discrepancy will need to be immediately addressed, since it may indicate potential disagreement with the subsequent reporting of any results. If no decision has been forthcoming in the paper so far, the selection arrived at should be noted and re-appraised when the results are finally presented.

Have the necessary assumptions associated with the stated or implied analysis procedure been met in a way that suggests the reader can have confidence in the results of the analysis?

Notice that I spoke above only of a "preliminary" decision being made as a result of any consultation with this chart. Each of the statistical procedures selected will carry with it a series of prerequisites or assumptions for its correct application. Thus, the reader now needs to move on to the second of the two concerns mentioned at the beginning of this section: have the necessary assumptions associated with any proposed procedure been met to permit adequate analysis to be made (i.e., in which the reader can have confidence)? Assuming the specific analysis has been identified in the text up to this point, we should expect the researcher to be aware of, and have adequately addressed, the assumptions associated with their choice. Nevertheless, the reader will need to think about whether this is indeed the case for each assumption, since few editorial authorities require them to be specifically addressed in the paper itself. Furthermore, the realisation of some of these prerequisites can only be conjectured at this point in the proceedings and confirmed later when we are in possession of the results.

I have used the word "assumptions" deliberately here. They are not "rules" for using these analyses. Some of these assumptions are of more or less consequence depending on the nature of the procedure; others must be addressed if subsequent results are to be presented with confidence. Furthermore, failure to meet a particular assumption does not mean that the researcher cannot process the data with *any* statistical procedure. As we shall see below, and depending on the assumption potentially violated, an alternative or less "powerful" test will often be available. What the reader will have to assess in the face of assumptions not met, and/or less powerful tests adopted, is the extent to which his or her confidence in the outcomes is thereby weakened.

The table[24] (see Appendix) describes the assumptions associated with key statistical tests or procedures. Independence of groups or observations refers, in general, to the fact that a score given to one case must not bias the score of another case. In the case of group bias, this assumption would be broken if, for example, information was thought to have been easily passed between groups. Similarly, we would want to be assured that the researcher checked to see that no member of one group was able to appear in another. This is a particular problem, for example, when dealing with large groups of subjects in similar sections in a school. If subjects are able to switch attendance between classes, a potential problem of this nature can arise. Clearly, any design that requires repeated measures to be taken from one group will not meet this requirement; happily, for many of the key descriptive or inferential statistical procedures we will see used, "repeated-measures" alternatives also exist. Similarly, independence of observation assumes that one individual score (or pair of scores in correlation) does not influence another. We can easily see how this assumption might be violated in research that calls for uninterrupted evaluation to be made of subject performance. Many readers might acknowledge how difficult it is to remain unbiased when scoring students' speaking performance one after another in an interview format: sometimes we might inadvertently judge the present student's performance by comparing him or her with the person we heard immediately before. In both these assumptions of independence, the reader will need to attend carefully to the information supplied by the researcher about data-collection in the "Procedures" section.

As can be seen from the table, normality is a key assumption in most of the commonly-used procedures for data analysis used in experimental and quasi-experimental research in our area. Initial information about this assumption will often need to be gleaned by the reader from the descriptions of the subjects and their background/learning context and the selection procedures themselves. Other, more specific, information should be available in the "Results" section. There are two questions for the reader to think about here. Firstly, the more powerful, parametric tests of analysis (see below) assume that the data themselves form a **normal distribution**: the mean and the standard deviation would be suitable statistics to describe the distribution. Secondly, we should be able to estimate a normal distribution in the population from which these subjects have

24. Adapted from Brown, J. D. 1992. Statistics as a foreign language: Part 2. *Tesol Quarterly*, 26, 4, pp. 629–664.

been drawn; in other words, that these are "representative" of the same population outside the "laboratory". This latter tends to be more difficult to judge, of course, but the reader might initially rely on common-sense judgements in their appraisal. For example, if we read that a researcher is interested in studying the way second-year students in University X in Country D go about learning new L2 (German) vocabulary, the population to which the results are directed is limited. Thus, we might feel that two groups of 20 subjects might well manage to be representative of the nature of the population the researcher has chosen for the study. Conversely, if we read that the researcher is interested in the behaviour of similar learners in the many other universities in Country D, two such groups (even if these were randomly selected, rather than in intact classes) are unlikely to be representative of the target population.

Our first, and less specific, appraisal of normality will come as a result of our reading about the numbers involved in the study. We might expect to see a normal population distribution formed when the number of subjects is large enough. "Large enough" is relative, of course, but the larger the sample size, the greater the possibility of meeting this assumption. In many of the studies in our area, where subjects are drawn from intact classes in educational institutions, sample sizes tend to be small. In such cases, normality in the data would be doubtful and the presentation of typical measures of central tendency and variability, such as the "mean" or "standard deviation", might not be immediately meaningful. More specific information on normality of scores can be drawn from these very measures in the "Results" section. Try to imagine the typical **bell-shaped curve** of the normal distribution; look for the mean and standard deviation (s.d.) figures in the descriptive statistics (hopefully) provided; finally, assume normality in the data if we roughly estimate that there is space in the curve for two or three of these standard deviation units on *either* side of the mean, and if there are apparently no extremely high or low scores obtained which might **skew** any potential curve (see below, "Results: the presentation and nature of findings"). Clear evidence of lack of normality in a distribution should be treated with extreme caution by the reader; it will make the interpretation of results difficult and seriously affect the conclusions that can be drawn from them. Again, if this assumption is potentially violated, an alternative less-powerful, procedure would normally be available (see below).

As the table reveals, the assumption of equal variance is one that applies specifically to those tests or procedures in which differences between groups (rather than correlations) are sought. This assumption is also very much linked to the measures obtained above from the distribution. Since, mathematically,

the variance is obtained by squaring the s.d. figure, the reader could obtain the variances of the groups involved and check that they are about equal. If they are seriously different, it would again suggest abnormality in the distribution, of course. Having said that, statisticians state that many of the more "popular" tests of group differences are strong enough to withstand violations of this assumption, as long as the group sizes are not highly dissimilar. Thus, the reader would do well to check up on the group sizes used in the study: problems are most likely to arise here when studies use small groups of highly uneven size. In obvious cases of inequality between groups, we would be looking to the researcher to have made the appropriate homogeneity of variance calculation and addressed the problem.

The assumption of **linearity** is one which can only be judged from the data obtained in the subsequent "Results" section of the paper. The "line" we will need to be looking for there will ideally be found by drawing a straight line through points of data marked on a **scatterplot**. Basically, if a sufficiently straight line can be "seen" through these points such linearity can be assumed. Once again, the scatterplot, or a description of it, may not be provided in the paper, so the reader will need to be particularly attentive to this potential problem when reading about studies in which linearity might be expected to fail. Imagine, for example, a study of the relationship between performance on a test of target-language phonetic transcription and time given up to study. At first, the transcription of what is heard into phonetic script is often difficult for many students. Little progress relative to time is demonstrated, and there will be a **negative correlation** (i.e., no linearity) between the two axes of "time given up to study" and "phonetic test". However, once they "get the hang of it" — and it can be an overnight change! — study time begins to pay off and relate more to successful outcomes on the test. The relationship between the two would then be expected to continue in a more linear fashion.

Related to this last prerequisite is that of (non)**multicolinearity**, since this assumption requires that the variables not be highly inter-correlated. To check on this, the reader will once again need to see tables of correlations in the "Results" section of the paper to check if there are any correlations of .85 or above (or equally high negative correlation) among a number of the independent variables. When this happens, it will be very difficult for the researcher to interpret which is relating in more or less fashion to the dependent variable. Finally, **homoscedasticity** is threatened when the scatterplot reveals anomalies between sets of data. The reader needs to consider if the data points from the two variables around the straight line are at approximately similar distances

apart all along the line. If not, this would tend to indicate that the amount of variability was not constant.

We should always keep in mind in our appraisal what claims the researcher wishes to make for their data, since the strict application of the specific assumption may change according to these needs. It is advisable to consult the specific manuals and reference books listed at the end of this book to understand what assumptions are considered fundamental to each statistical procedure and which are more flexible. We also need to be informed about the alternatives facing the researcher and what could or should have been done when one or more key assumptions were not met. Those relevant to the papers studied here will be described and discussed in the relevant workbook sections below.

For example, as can be seen from the table, the assumption of normality is one which is a *sine qua non* in the majority of the data analyses here. A researcher should be seen to have aimed for the most "powerful" test of their data. "Power" relates directly to an analysis that gives him or her more confidence in making any claims about eventual outcomes. Nevertheless, statisticians differ in their opinions about how comparatively powerful certain groups of tests really are. Normally-distributed data allow the researcher to choose from the ample range of powerful *parametric* analyses. These tests are more powerful because characteristics of the normal distribution are well-known and have been calculated for. The argument is that *non-parametric* tests, by putting scores into a ranked order, only measure the variability in scores indirectly and therefore do not take into account more information about the numerical differences between these and, consequently, the experimental conditions that produced them. Parametric tests are thought to be more sensitive procedures. This said, non-parametric tests *can* be used with interval or score data which the researcher — for some stated reason — did not feel confident about using in a more powerful procedure. Information about data would, however, be lost in opting for such an alternative. If we read that a researcher fails to meet certain key assumptions and opts for a non-parametric alternative, rather than criticise the statistical procedure as weaker, we need to consider the reasons for, and consequences of, such a decision, and the implications for the data collected. In our appraisal of papers in the workbook, the assumption will be that the researcher might initially aim for the most powerful data analysis.

Finally, we have noted throughout this book how our research contexts in the field of second language learning inevitably restrict the kind of conclusions we can obtain from findings. When our subjects are drawn from small intact classes, for example, and class/subject selection cannot be done randomly, it will

be impossible to meet some of the assumptions required to use a statistical procedure for inferential purposes. In such cases, these statistical analyses can still be used to give us confidence in the descriptions of the data we are to be presented with, but *not* in the possibilities of extending these descriptions to a larger population. Whatever the eventual function intended, however, remember that the researcher should be aware of the need to have met the assumptions that lie behind the appropriate use of these procedures as far as possible. In circumstances of serious violation, the reader's confidence in the eventual findings is inevitably weakened. If one key assumption is ignored, and the researcher goes on to use the specific analysis normally, there will most likely be distortion of the significance levels reported. Since the end result could easily be meaningless data, the reader would then be looking to the researcher to qualify the results presented so that we can decide whether the end (i.e., the kind and quality of data obtained) justified the means (i.e., the use of a statistical procedure for which some important assumption was not met).

3. Results
The presentation and nature of findings

Does your initial reading of this section suggest that enough data have been provided so as to have adequately responded to the research questions or hypotheses previously put forward?

Although in many papers the results and discussion sections will be compressed into one, they will require different approaches to appraisal and so will be separated for present purposes. This said, our reading of any results will inevitably be linked to our interpretation of these, and it will be useful for us to start considering what any data on the page are telling us — ideally before discovering the researcher's own interpretation of these. The immediate aim for the researcher now is twofold: he or she must report what happened in a readable and easily interpretable form and describe those findings in a succinct way, prior to more profound discussion of their perceived importance in the subsequent "Discussion and Conclusions" sections. Most journals, for reasons of space, require brief summaries of results here. However, it is to be hoped that these include enough information about outcomes to be seen to have adequately responded to the research questions or hypotheses previously proposed. As readers, we will need to recall our appraisal of the research questions or hypotheses and check that — at least on a first reading — what was required by these has indeed been provided in terms of the data in this section. Obviously, the findings presented here will also need to be consistent with the method, procedures, and selected data analysis described earlier. Consequently, once again, our appraisal of the nature of outcomes needs principally to be based on our accumulated knowledge and evaluation of the study so far.

What tables or graphical displays of results are provided, and what do you understand from the data displayed? Are there any data that you feel might have usefully been added to the information provided here?

In the kinds of study that concern us in this book, the initial presentation of findings will often be by means of tables that help summarise the quantifiable findings in some easily-read form. Again, we will find variation in the amount of detail editors recommend to be included in such graphic presentations, but many agree with the *APA Publication Manual*[1] that such graphics should supplement (rather than duplicate) information in the text.

Generally, these presentations highlight the most significant information and results for the reader. As we shall see, however, the reader will again need to be attentive to both kinds of presentation of results (i.e., graphical and textual), since it is sometimes in this very data reduction that interesting tendencies go unobserved or weaknesses are concealed. Equally, we will need to be alert to any data which we feel might be missing or which might have provided further insight into what actually happened.

A useful initial awareness-raising procedure, therefore, is for the reader to locate these displays and try to interpret what he or she is being told therein *prior to* reading any reporting or interpretation of the same by the researcher in the main body of the text. This allows us to come to our own preliminary conclusions about the data, and it can also help clarify any previous doubts about the data analysis involved. As a result, we might be able to foresee points we would then like to see followed up or further explained in the text itself. Brown (1992) provides the basis for a useful check-list procedure that can be used as a first strategy in interpretation:[2]

Examine the table and, in the light of the statistical procedure used, confirm

a. what results it purports to show and the relevance of these to the study and analysis procedures proposed.
b. what the column and row labels and/or sub-labels stand for in the study. Check that these labels correspond to the groupings and/or variables envisaged in the "Method and Procedures" section.
c. what any statistical abbreviations refer to and their relevance to the analysis procedure.
d. and decide on our initial reactions to any significant/non-significant outcomes described.

---

1. See, in particular, sections 3.74 and 3.86 in the *Manual.*

2. Adapted from Brown, J. D. 1992. Statistics as a foreign language: Part 2. *Tesol Quarterly*, 26, 4, p. 651.

Similarly, the reader needs to be able to interpret any graphs and figures that have been used as a means of providing concise information about results. As we saw in the last section, for example, a number of assumptions associated with specific descriptive or inferential statistical procedures require information about the distribution and **dispersion** of data, the linearity, or homoscedasticity, much of which can be presented graphically through **bar graphs**, line graphs, **histograms**, and scatter plots. Although the way all this is presented varies from paper to paper, the reader needs once again to begin (or continue) his or her appreciation of results by responding to what these figures reveal, rather than be told about them initially from the (potentially subjective) viewpoint of the person who produced them. For example, bar graphs and **polygons** (line drawings) are often useful initial insights into the descriptive statistics behind a particular study. The reader should look at the ways axes are labelled on these figures and begin to think about the implications behind the graphical display produced. In the case of polygons, for example, we might be being shown frequencies in relation to one another. Such displays may reveal results that seem not to belong in the data, or that seem to suggest that these were not as expected across the sample. In this case, the reader would want to start thinking about why this may have been so and the effect this observation might have on any further statistical analyses yet to be carried out. All this, of course, prior to confirming that the researcher has also noticed and addressed the question in the text.

Having read and come to our own initial conclusions about what is understood therein, we might now look to the researcher to direct our attention to those findings perceived as being most important. The *APA Publication Manual* suggests particular focus should also be on any results that are inconsistent or "run counter to the hypothesis" (p. 15).

> **What information is provided by any descriptive statistics about the distribution of data?**
>
> **Have the data been scored using the unit measurement predicted earlier and/or has any appropriate data conversion taken place?**

In the kind of studies we have been concentrating on in this book, we could be looking for some information in graphical and/or textual form concerning the descriptive statistics for a given relationship or effect. Such results are to be considered as basic identification data, wherein the reader can obtain some initial idea about the distribution of data in the sample studied: how results

were spread out across these subjects, how often a certain observation occurred, or how typical these observations were amongst these subjects. Throughout most of the analyses that follow, we will need to be looking out for crucial indicators in these statistics, some of which have already been mentioned as providing us with essential information to decide on whether specific assumptions have been met appropriately.

Frequencies (f) are reported when some kind of counting of certain phenomena has taken place. These often begin as simple frequency counts where each subject is categorised (e.g., male/female, NS/NNS) or categorises themselves (e.g., yes/no answers given), and move to frequency-based *scores*, in which each subject might be awarded a percentage mark for performance based on the number of correct or acceptable answers. The former kind of results are often presented in graphic form through bar charts or graphs and descriptions offered, while scores might then be summarised and subjected to closer scrutiny and comparison through measures of central tendency and variability (see below). It is often impractical to provide such detail with large samples, but the reader should remember that individual **frequency distributions** (of certain scores on a test within a class, for example) can provide important insights into particular tendencies *within* a group before any comparison is made with other groups. Such data are usually lost or unseen when only collective or mean data are offered.

Frequencies as raw data are of a nominal **non-continuous** nature, but researchers will often be seen to convert them into **continuous**, score data such as percentages or rates in order to be able to use these in more powerful parametric statistical procedures. As I have mentioned earlier, any data conversion needs to be fully explained or justified and the reader satisfied that the converted data meet any other assumptions applied in the statistical procedure in which they are then used. Conversion into percentages is a case in point: the researcher (and the reader) will need to think about other factors which go to make up a percentage total before coming to any conclusions about what appears to amount to just a simple score. Ideally, we would want to be informed of both the frequency *and* the converted per cent score since the raw frequencies might be small even if the per cent score looks impressive. For example, if a question was answered correctly by 5 subjects in class A (n = 5) and this same question was answered correctly by 20 subjects in class B (n = 40), the percentage scores would be recorded as 100 % who got the correct answer in class A, and 50% in class B. As readers, we would want to look beyond the initially impressive 100% score-line to see how meaningful such a score really is in

terms of the original raw frequency when placed in its initial context. Secondly, if the original groups contributing these converted frequency tallies were unequal to start with, it may well be that total percentage scores of correct answers from both groups conceal the fact that the larger (or more able) group contributed more to the final percentage (or had more opportunity to do so).

Continuous data may also be achieved by the conversion of raw frequencies to rates. This is often used in studies in which the frequency of some textual feature needs to be tallied. Thus, it would make more sense in such a study to report, say, the number of times a particular word occurred "per 100 words" than simply report a frequency tally. As with percentage scores, the conversion means that data can now be compared, rather than only described. While such conversion does not need to be explained in such circumstances, the reader might want to consider how logical, acceptable, or appropriate the final unit (i.e., "per *x* quantity") appears to be in the context of the study. For example, research that used textual features as its "subjects" and extracted data from a corpus might well report a standard unit of "per 1000 words", but such a rate would perhaps look rather excessive when data are drawn from 5-minute recordings of L2 beginners' descriptions of their daily lives. A unit claimed to be "per 100 sentences", for example, would require an operational definition of "sentence" in order to facilitate replication; in the same way, such a unit may seem inappropriate in a study where the object was to report the number of times subjects corrected their pronunciation of a word.

Perhaps the most useful descriptive indicators of typical group behaviour that we should be looking out for are the measures of central tendency and variability. Once again, these measures of basic differences between the groups may often go unreported in studies where they are then to be used in specific formulae for inter-group comparison and the testing of significance. "Significant" differences and relationships between groups and variables will need to be appraised in the light of the amount of difference or relationship there was in the raw data. Not describing such crucial data is unfortunate since, in the majority of cases, the information is extremely useful to help the reader understand better the way individuals within each group performed on a data-gathering measure, the distribution obtained thereby, and, therefore, the appropriateness of the procedures eventually undertaken to compare the groups. Any decision made implicitly or explicitly here will have implications for both data analysis procedures and the descriptive/inferential use that can be made of any outcomes arising from such procedures. It will also indicate a

certain predisposition on the part of the researcher towards the normality of the data he or she has obtained.

Of the three measures of central tendency, the **mean** is by far the most common in studies manipulating interval/continuous scored data. This is understandable in the sense that it is the measure most used in powerful parametric statistical procedures because of its stability and comprehensiveness. Notwithstanding such popularity, the consumer of the research will need carefully to consider the appropriateness of the choice made in the light of what we have been told. The mean is considered to be a comprehensive measure because it takes into account each and every score obtained in the group. No information is lost, and so any scores which somehow do not seem to fit in with the rest of the group's performance are also included in the calculation. This makes it a measure that is highly sensitive to extreme scores (also known as "outliers"). Thus, a couple of very high- or very low-scoring subjects in a relatively small group could displace this average measure to the left or the right of the middle ground and seriously affect the distribution obtained. When the distribution is abnormal, it is unlikely that the mean would be a safe measure of group tendency. As we mentioned before, for reasons of space, individual performance indicators are often subsumed in collective data. Yet there are times when such figures can determine whether any subsequent analysis and findings are valid or not. When outliers are identified, the researcher will need to decide whether to eliminate these data in order to use the mean (justifying their elimination and, perhaps, studying their uniqueness in another section of the research), or leave them in and use a less "demanding" measure of central tendency such as the median. Logic should tell the reader to be more wary of groups comprising small intact classes. In such contexts, (lack of random) selection procedures, together with the few subjects involved, may well skew the normal distribution so often assumed in the most popular descriptive or inferential statistical procedures. Going on to use the "all-embracing" mean as a measure of central tendency in such circumstances would be unwise.

The other two measures of central tendency are the **mode** and the **median**. The former is the score most recorded within a data set. There are a number of problems associated with its sensitivity and comprehensiveness. Most of these are logical consequences of using the most often obtained score. For example, if there happens to have been no one repeated score in a data set, there is no mode either! Also, one might imagine a disaster scenario wherein a particularly difficult test might produce a high frequency of zero scores in a group. In this case, one might be faced with using a mode that will hardly be useful for any

inferential statistical analysis that follows (although a series of 0 scores might be of interest in any subsequent descriptive statement). Having said all this, knowing the most frequently-obtained score in a group might reveal interesting tendencies if two very different scores were found to be often repeated in a data set. This **bi-modal** score distribution would be further evidence of abnormality in the group, since the suggestion would be that two different "groups" (i.e., with two distinct tendencies) seemed to have been formed within one group. The median is "in the middle" of a set of scores: fifty per cent of the scores fall above it and fifty per cent below. In this sense, it is less sensitive to extreme, outlying scores and would be the natural choice when there were problems perceived with using the mean. It too has its problems, since the calculation as to where exactly the half-way point between data lies is not always so clear-cut. Nevertheless, it remains a useful alternative to the mean when the numbers of subjects involved are small and/or where the researcher has reason to believe that he or she will not obtain a normal distribution of scores from the group data. For this reason, the median is mostly found in non-parametric inferential statistical procedures, which do not make any strong assumptions about the normality of the data obtained and use rank-order as the basis for calculation.

Like the measures of central tendency, two of the three measures of variability available differ in their comprehensiveness. The most common measures are the **range** and standard deviation (*s* or *s.d.*), both of which report how homogenous subjects are in their reported behaviours. As far as appraisal is concerned, the reader will need to be attentive to the kind and amount of information reported in one or the other measure. The range is probably the crudest and most unstable of measures of variability, since it takes into account only the difference between the highest and the lowest score obtained. Thus, only two scores enter in the calculation. Any intermediate scores are disregarded for the purposes of the calculation, which means that no information will thereby be available on how the rest of the group performed relative to the measure of central tendency. Furthermore, it too is alarmingly sensitive to extreme scores: imagine how one lowest score of 0% or highest score of 100% in a group might alter the range completely! Of the two remaining measures, the standard deviation is probably the most useful and frequently-encountered measure of all (the variance equals the square of the standard deviation and is important in the calculation of various inferential statistics).

Both these measures have the advantage over the range of taking every score into account. As readers, we need to understand what the figure reported is telling us about the data and then use it to appraise what we are told (or not told)

about the variability in the sample. The s.d. figure firstly informs us about how scores are spread out away from the mean in both directions. Thus, a high s.d. would indicate that the data are widely dispersed around that figure and that subjects did not perform very uniformly on the data-gathering measure. Secondly, once again there will be important implications for the normality of the data being used and, in consequence, for any analysis of these same data. **Standard deviation** may be thought of like a ruler, dividing off equal sections from the central point of the distribution. Statisticians have calculated the percentage of scores that occur in each of these sections, if the distribution is normal. So, for example, by calculating if there is "room" for at least two reported s.d. measures on either side of the mean (given the total score possible for the test), the reader will be able roughly to estimate the normality of the distribution, without actually seeing any descriptive polygon of these results. Similarly, if there appears to be room for more than three of these s.d. values above the mean, it would indicate that over 50% of the scores were above that point. This, too, would suggest some abnormality in the distribution and a larger than "normal" amount of higher scores on the measure. To take a concrete example, if we read that the total score available on a test was 50, that the mean was 40, and the s.d. 10, we should note irregularity because there would only be room for one s.d. above the mean. As a result of any apparent abnormality, the reader would need to be particularly alert to the subsequent (and probably unjustifiable) use of any analysis procedure that required normality as one of its assumptions.

Thus, in any descriptive statistics of scored data, it is important for the reader to be on the lookout for the reporting of *both* measures of central tendency and variability. The reason for this is that both offer us complementary pieces of information about the data. For example, it is quite possible for a researcher to report the same means across two sets of data collected. Conversely, we cannot then assume that the groups were the same. The only way of knowing if the groups really were identical would be to report a measure of variability that took into account every score and demonstrated if the distribution of scores around that same mean was equally comparable. Finally, even when normality can be assumed from what we have said so far, we would still be wise to check on the normality of the distribution curve itself. Such ***kurtosis*** is rarely reported in these circumstances, although computer programs can easily calculate it, but it provides useful information on the shape of the peak in the normal distribution curve. If we think of the shape of a very sharply-peaked "normal" curve, for example, we will understand that this describes a high

frequency of results bunched closely around the mean line; the s.d. figure would most likely be small, confirming the minimal dispersion from the mean. Although the resulting curve is symmetrical, it may reveal a certain "abnormality". With the normal distribution threatened again, we might expect problems in any subsequent statistical analysis.

i. **What, if any, specific statistical operations or calculations were carried out on these data, and does this seem to have been carried out appropriately?**

ii. **In the light of what you have read in this section, do you wish to amend or add to your previous appraisal of the assumptions met for this procedure?**

In many papers, providing a description of the data through descriptive statistics may be the sole purpose of the research. However, these give us part of the picture, only telling the researcher and the reader about the tendencies in this particular sample and, where relevant, the differences between samples or the relationships between variables within a sample. The problem is that the data so described are based on samples selected and assigned to specific groups in the study. Group characteristics will inevitably vary greatly from one group to the next throughout the many L2 learning settings around the world. For this reason, we need to be cautious when reading conclusions based exclusively on calculating such descriptive statistics for one particular sample. The only way possible of checking on the reliability (and eventual generalisability) of such conclusions would be to replicate the study in many contexts and with a number of different samples. Such a method would prove a fairly reliable way of confirming conclusions in the long run, but is an extremely expensive and inefficient way of conducting science, since it relies only on the accumulation of knowledge rather than the considered interpretation of what we already know. For this reason, many researchers turn to inferential statistics when they wish to look more profoundly into the data obtained and, perhaps, say something about similar populations in similar settings.

Although the set of specific statistical operations we are about to discuss has this denomination of "inferential", it should be remembered that the same procedures can be used (and the significance level reported) to establish confidence only in the *description* of the outcomes. We would still be looking to check that the relevant assumptions behind each test or procedure had been met, of course, but there is no attempt or desire on the part of the researcher to go one

step further and generalise beyond this description to other settings. In this case, the analysis is being used to give the researcher confidence that the calculated difference between groups or relationship between variables is existent for that one research context. As we have seen in many sections of the papers appraised so far, particular care is required on the part of the cautious reader here, for there can be a temptation to take the step towards generalisation without the minimum support of a research design which permits such a move (i.e., which meets the principal threats both to internal and external validity).

The common factor behind all of these procedures is the search for differences or relationships that show a significant enough level of statistical probability for the researcher to suppose that their occurrence is not due to chance factors alone. It is clearly beyond the scope of this book to describe each and every descriptive or inferential statistical procedure used in our field of study. In what follows, I have attempted to alert the reader to what needs to be considered when appraising results from some of the most common statistical procedures used in the field of second-language learning. I have assumed that readers are already familiar with the basic objectives, method, and function of each procedure; my focus here is on appraising the consequences for interpretation of using such a procedure to analyse the data. Space also dictates the specific variety of test (i.e., parametric/non-parametric) and design (i.e., between-groups/repeated measures) that can be discussed. Although our appraisal of results will be very similar in most cases, readers are encouraged to make use of the manuals listed at the end of this book for more detailed information about other versions of these tests. Furthermore, while the principal criterion for selection has been their widespread use in the field, these procedures (with the possible exception of regression) have been selected also because they are appropriate to the kind of research context in which many readers of this book will find themselves. Thus, I have also assumed that readers will need to be more familiar with procedures that admit relatively small numbers of subjects or observations and can be carried out without the obligatory use of statistics software. Readers are again referred to the references in the "Further reading" list for an explanation of more complex procedures.

Correlation

When the relationships between variables are being described, researchers commonly use correlation procedures to obtain the relevant correlation

coefficient, which will normally be quoted as "*r*", "r_{pbi}", or "ρ" (rho), depending on the specific procedure used. We can expect to see reported a coefficient that ranges from −1.00 (the strong relationship between variables moves in opposite directions) to 1.00 (the relationship moves in the same direction). It should also be remembered that the sign "+" or "−" would only indicate the direction of the relationship. Thus, a correlation coefficient of −.50 would show us a stronger (more predictable) relationship in the negative direction than one of +.30.

In practice, it would be unusual to see many correlations of 0.00. The problem for the critical reader is that even "non-correlations" such as these can turn out to be statistically significant! This is because some degree of correlation will almost certainly always be present through chance factors between two sets of score or numerical data. We will need to pay particular attention to the number of subjects involved, since reported significance (and our appraisal of this) will always depend largely on sample size. So, for example, although a correlation of .65 may not be seen to be statistically significant with a small sample group of subjects, it may well have turned out to be so had the researcher involved larger numbers in the study (see below, "**effect size**"). As so often in our appraisal of these studies, the focus will be on just how significant is "significant" given the relationship being tested and the context of that same relationship.

The tabular presentation of correlation results of a number of variables will often be through some form of correlation matrix:

Variables	**Reading**	**Writing**	**Speaking**	**Listening**
Reading	1.000	.425*	.315	.754*
Writing	.425*	1.000	.442*	.663*
Speaking	.315	.442*	1.000	.851*
Listening	.754*	.663*	.851*	1.000

* $p < .01$, $df = 62$

In this fictitious example of a **Pearson correlation**, we are shown how four independent variables (sub-tests of L2 proficiency) were seen to correlate. Since the relationships described are symmetrical, tables will normally be presented with only the top half filled in. Working down and across from each side of the matrix, we can see the origin and outcome of each correlation made. The researcher also presents the significance level ("$p < .01$") of each correlation,

ideally based on the preferred "alpha" (α) level, which many editorial authorities require to be stated before any calculations are presented (see below).

Researchers will want to have used a statistical test to give them support for their findings, in order to be sure that any measured difference or relationship between groups was not due to chance alone. These "*p*" figures refer to the *p*robability level calculated. In this case, ($p < .01$), the researcher is stating that there is less than 1 per cent probability that an observed relationship this great would have been down to chance alone. It is also useful for the reader to pause and think a little about both the "p" and "alpha" statistics offered. The "alpha level of significance" is the true cut-off point for judging what is likely or unlikely to be due just to the chances of sampling in a study. Ideally, we would be looking for the researcher to report this *before* any calculations are made. He or she might have had some stated or implied reason (perhaps based on previous findings or constraints on subject selection) to select a higher or lower (alpha) level of probability for any analysis. By previously opting for an acceptable level of probability, the researcher is seen to have avoided the temptation to have made the eventual probability reported sound as if it were what he or she was expecting anyway. On the other hand, we may often come across studies that only report the level obtained *after* calculations, as "p". In such circumstances, the reader would need to assess the appropriateness of the level now reported in the context of what has been read. We would need to be particularly wary in such circumstances of any suggestions that a particularly low level of probability obtained (i.e., $p < .00001$) was somehow "better" or of "greater importance" than if the result had been significant at a higher level (i.e., $p < .05$) (see below, "t-test").

The "*df*" figure (degrees of freedom) is an adjustment that the relevant formula takes account of in calculating observed statistics. Again, editors may or may not request this measure in data reporting, but it will help us to check on the probability level obtained and the significance of any results in statistical tables, assuming the researcher had calculated this by hand. These adjustments vary according to the procedure being used. In correlation, the *df* should have been computed as the number of pairs of results less two. In this case, for example, 64 subjects took each sub-test; therefore, the *df* value is 64–2. In the table provided for the Pearson-product correlation procedure there is no line that specifies 62 as the *df*, so the researcher would normally have opted for the more conservative 60. The intersection of the *df* row and the p value shows the critical value needed (equal or greater than this value in this case) to reject the null hypothesis. Nowadays — with the help of complex computer software —

it is unlikely that researchers would have committed errors in the actual calculation of results; however, it may be useful practice to check in the tables provided at the end of this book — as far as possible — on the correctness of critical values and significance levels, given the *df* involved.

In our appraisal and interpretation of the correlation coefficient, our critical attention should initially be drawn to two measures: the significance level reported and its importance in this context and the reported (or calculated by the reader) shared variance between the variables. As we shall see below, this will be the first of several occasions in our interpretation of statistical procedures when we will need to take care not to be dazzled by apparently impressive and highly-significant relationships. Statistical tables of correlations (or computer software) will give the researcher the information required about whether a certain relationship is to be reported as significant or not, but both the researcher and the reader will then need to interpret that outcome based on the context and content of the particular study and the results of any previous work on the same or similar sets of variables. Let us imagine a study where a researcher studies the relationship between L2 writing proficiency and L2 grammatical knowledge. We might expect the relationship to be a high one in many language learning contexts. Perhaps previous studies have reported this strength as significant at an *r* of 0.80 or more. If, after analysing their data, the relationship turned out to be "only" 0.60, we might still expect the researcher to address the reasons for the apparently "weaker" result, even though it turned out to be significant from the tables. Conversely, if a researcher is studying a relationship between two variables that would not normally be hypothesised to be highly related and/or for which little previous research exists, a weak (but also significant) relationship may be a noteworthy result and just as interesting to the field.

We should also bear in mind that a reported coefficient only tells us how well the two variables "fit" together. To understand the relationship better, it is more useful to see how far this relationship actually indicates that the two variables overlap or are providing similar information. An accompanying initial scatterplot reproduces graphically the direction of the relationship between two variables as plotted points on a graph around an imaginary straight line (i.e., representing what would have been a perfect correlation between the two variables). It is a useful additional aid to help the reader see what is going on and can often provide information that is hidden from view in the concluding *r* statistic. If such a scatterplot is presented in the text, the reader might wish to check on the apparent positive (bottom left to top right slope) or negative (top

left to bottom right slope) direction of the line through the plotted points and the general shape described. In particular, we would want to focus on any deviation away from the straight line and check to see if the graph looked linear or **curvilinear**. The strength of any correlation will be graphically represented in the way the plotted points cluster near the imaginary straight line; the closer these are to that line, the stronger the relationship between the variables. Similarly, if no apparent line can be drawn through the points because these are so dispersed around the scatterplot, there is little or no real relationship between the variables and the causes would need to be addressed.

In the Pearson example, such shared variance or **strength of relationship** is easily calculated simply by squaring *r*. Thus, a reported *r* of .52 significant at $p < .05$ would show a shared variance in our study of .27, suggesting that 27% of the variance in L2 writing proficiency is accounted for by L2 grammatical knowledge. While this appears to be a fairly strong correlation, just how meaningful that overlap really is again comes down to interpretation based on what we already know. Such a percentage may not seem to indicate the kind of overlap we might have expected in the circumstances. Much depends on the objectives of the researcher: if he or she had previously hypothesised that two measures should be testing the same thing, such an outcome would be surprising. Conversely, he or she might be content with even a small, but significant, shared variation if previous literature had suggested that there would be no significant relationship between variables, or that perhaps this relationship would exist in an opposite direction. Whatever the outcome, we might want to make a mental note to see this point addressed in any upcoming discussion.

This interpretation might also need to address what is classified as "error" in such an overlap (i.e., the remaining 73% of variance), and which is presumably largely due to variables other than those studied. Such "error" cannot just be ignored and the focus of scientific attention turned onto the calculated strength of relationship. This would be particularly important in cases where the researcher had hoped to show that the two variables basically measured the same thing. If the correlation is quite weak (and the more error is involved in the relationship), the less confidence the researcher will be able to have in any predictions of the outcomes from one variable based on the other. One formal way of addressing error, particularly if the researcher is interested in using a supposed correlation for the purposes of prediction, would be to obtain a measure of this error (known as the "standard error of estimate"), often quoted as "**SEE**". This figure acts in a similar way to the s.d. measure, in that it shows the researcher the amount of dispersion around that straight line drawn

through the scatterplot of the relationship (known as the **regression line**). This measure is useful on certain occasions for further interpretation of the overlap. It will tell us how much error may occur if we go on to use the correlation coefficient obtained to predict scores on one variable based on scores from the second variable. The larger this figure is, the greater the chance of error in predicting a score. As with the s.d. measure, there is no recognised cut-off point of acceptance. For the purposes of our appraisal, we simply need to recognise that the value gives us information about how wrong the researcher might be in predicting scores from the coefficient.

A cursory glance at our flow chart shows that the Pearson *r* is the only correlation procedure that works with interval data from both variables. Other correlation procedures require data that are ordinal or rank-ordered, while others can even be applied to nominal data. Of the non-parametric tests, one of the most popular is the **Spearman** *rho* (ρ), used to compare two sets of ranks to estimate their level of equivalence. If the original data have been obtained in interval form, this can be converted (important information is, of course, lost thereby) to ranks and this test applied. A researcher might deem this useful in cases where the **raw score** data somehow seems to make more sense once it is ranked, for example when individual examiners are seen constantly to respond differently to the same marking scale. The computation of ρ includes an operation of squaring differences, and so we cannot talk about strength of relationship in this case as ρ^2 . Rather, the significance of ρ will be confirmed in the tables, although, once again, the strength of the correlation gives us more information than the significance as such. We should also be aware that the Spearman *rho* does not respond well in situations where there are a fair number of equal ranks in the raw data. In such cases, the researcher would be advised to opt for Kendall's *tau* (τ).

There are a number of other issues affecting correlation that need to be remembered as we digest the results (and discussion) of any analysis. By far the most important of these focuses our attention on the nature of the correlation itself. There is often a dangerous temptation to read much more into an observed co-relationship of variables; however, correlation itself cannot be used to suggest that the relationship between variables is causal, and to do so would indicate serious misunderstanding of the nature of such an analysis on the part of the researcher. Correlation data can be viewed in the light of cause-effect hypotheses in that they may show up areas that might usefully be subjected to hypothesis testing in further research. For example, if a high positive correlation is obtained, the arguments for a cause-effect hypothesis might be strengthened.

But recording a significant correlation, low or high, does not demonstrate that one variable "causes" the other, or that one variable "affects" another. It only establishes a significant relationship — positive or negative — between them.

Secondly, there are a number of factors that relate to the distribution of the data obtained. As on previous occasions, when group data are simply plugged into formulae and run through the computer, this information may be lost from view to the reader (and the researcher), although descriptive statistics reporting the dispersion of data and measures of central tendency should provide valuable clues. Yet one or more of the following is sufficient seriously to distort any correlation analysis findings. **Outlying data** from one variable, (i.e., a result that does not seem to fit in with the rest of the data) may affect a coefficient. Likewise, we might also want to be sure that the data obtained do not bunch at extreme points along the continuum of scores, leaving large areas of the scale without any scores. This **accumulation of high or low scores** should also be evident from the descriptive statistics given. In both cases, data need to be distributed throughout the range of possible scores as far as possible; otherwise the assumption of normality cannot be met and no (parametric) correlation analysis appropriately carried out. This question of distribution of data also extends to the subjects themselves. The researcher should have taken care that the sample from which data have been obtained has the necessary characteristics to guarantee a full range of scores in the first place. For example, if a researcher has decided to correlate L2 reading ability with age but uses only subjects between the ages of 12 and 15, one of the variables will inevitably contain data from a very limited database. If **initial restrictions on the range of data** have been imposed by the subject sample itself, the correlation coefficient might turn out to be lower than it might have been if more wide-ranging data had been obtained.

Finally, the value of correlation — like any statistical procedure — ultimately depends on the quality of the data used to measure it. Ideally, therefore, the **original data would have been tested for reliability**. If either of the sets of data to be correlated has not been reliably measured, correlation can be affected in some way. Particular care should be taken in studies that claim to correlate a specific test with another L2 variable. We may find that the test — particularly if it is a commonly-used example — comes with its own reliability coefficient (which should have been quoted in the paper), but that the variable with which it is to be correlated has not been measured so carefully. If little or no formal information about reliability is provided, and the variable is then related to a test where reliability *has* been precisely calculated, the end result is unlikely to

be one in which we could have great confidence. On the other hand, if the reliability was calculated and the researcher saw that this was clearly dissimilar in the two (or more) variables about to be correlated, an attenuation formula[3] could be applied to the reported reliability, which would suitably adjust the correlation.

It should also go without saying that, in any appraisal of results, the reader should think back to their mental notes made in response to the assumptions apparently met (or not met) in the previous section of the paper. Now is the time to return to these thoughts and see whether the information provided in this section of the paper suggests we add to, or modify, those observations.

Regression

Regression allows us to go further than correlation and use the established linearity of the relationship (as shown in the straightness of the regression line) between the two variables to predict scores on the dependent variable from one or more independent variable(s). If the researcher claims to analyse the data using *simple* or *linear* **regression**, the prediction is of only one variable based on the scores from the second. In the more wide-ranging *multiple* **regression**, the researcher has the opportunity to ponder a number of independent variables and report, not only on how each predicts performance on the dependent variable, but also which combination of these variables might better predict outcomes on the dependent variable.

It follows that our attention might first be drawn in such procedures to the correlation coefficient itself. If the reported r is close to 1 (i.e., the perfect correlation), we can have confidence in any prediction of performance on the dependent variable. The opposite is the case if there is no correlation (i.e., the r is around 0). If the error in prediction (often referred to as "residual" in reporting results) is likely to be great — in other words, if the correlation is weak — it will probably be best to turn to the mean score itself as the best (albeit informal) prediction of results on the dependent variable. Simple regression adds little to correlation, in fact, unless one wants to predict individual scores on one variable from another. If we see that the correlation was not

3. A worked example of such a correction can be found in Hatch, E., and Lazaraton, A. 1991, *The Research Manual.* New York: Newbury House Publishers (pp. 444–445).

strong to start off with, but that regression has still been used to predict results, we might expect to see the researcher report the formal SEE value (see above), so that the reader can also appreciate how wrong the researcher is likely to be in predicting scores for one test by using scores from another one. Analogous to the s.d. statistic, the SEE tells us about the dispersion of scores away from the regression line; thus, the greater this statistic, the more marked is the spread away from the line — and the more likely error will be made when predicting scores on one variable from those on another.

Any results from linear regression will have used descriptive statistics that we should see reported (even if they have been subsumed in the final calculation): the mean score on test A, the mean score on test B, and the calculation of slope of the regression line between the two variables (usually quoted as *b*).

Multiple regression is a more versatile procedure and provides a subtle description of the predictive effect of a number of independent variables on the dependent variable. It also has the advantage of showing the researcher which is having the strongest independent effect among several that may all separately have significant correlations with the dependent variable. As we read through the results of such regression, the reader needs to remember (not for the first time!) what is understood to be behind any figures reported. In other words, regression "starts where correlation left off", and so we are being asked to assume that any correlation values plugged in to the regression formula are accurate. "Accuracy" here includes attention to key assumptions for correlation procedures, particularly when regression is being used with an eye to generalising from any outcomes. As can be seen from the table, multicolinearity threatens multiple regression, and we will want to pay attention to the correlation coefficients to make sure these are not similarly high. As with many other values plugged into formulae and perhaps thereby concealed from view, it would be as well to see the value of the coefficient and be assured that any combination of variables to be assessed for prediction were, individually, based on reliable measures from the outset (see above, "Correlation") and, if necessary, corrected for attenuation. Also, it follows logically from what was said earlier about sample size in correlation that — if more variables are going to be involved in predictive procedures — more subjects are going to be needed. Most statisticians agree that multiple regression is not applicable to small samples and recommend at least 30 subjects for each independent variable involved in the calculation.

As always, attention will need to be paid to any tabular presentation of results, particularly since a number of relative contributions to variable predictions will

be displayed. Results tables will therefore normally inform about the comparative predictive weight of each independent variable either separately (entering the variable data all at once) or as each one is added (entering data in a "stepwise" fashion) on the dependent variable and the statistical significance level of the outcome, as in the following fictitious (and simplified) example of the predictive power of performance on an L2 spoken production test. In this example, data have been entered in a stepwise fashion:

Variables	*r*	R^2	Change in R^2
Error correction	.67	.45*	.45
Intonation	.54	.46*	.01
Vocabulary	.45	.48*	.02
Content	.36	.48*	.00

* $p < .05$

In this example, imagine that the four variables have been found to provide a statistically significant contribution to the dependent variable. The table then has to show the reader the quantity of variance in the dependent variable scores (in this case, the test of L2 spoken production) that is explained by each independent variable (R^2). Immediately, we can see that "error correction" is a good **predictor** of performance on the test because it shares 45% of the variance. The "Intonation" independent variable has an R^2 of .46, showing that — once added to the "Error correction" variable — the combined variance is 46%. The final column (Change in R^2) describes the relative gain in predictive power obtained. Thus, "Intonation" only accounted for an additional 1% in prediction. If the next variable, "Vocabulary", is plugged in to the calculation, the three now make up 48% of the variance and 2% more predictive "value" is added. "Content", although it was significant in its contribution, does nothing to improve on this cumulative predictive value. As a group, these four independent variables explain nearly 50% of the variance in L2 spoken proficiency on this measure. Furthermore, the researcher has tested to see if the R^2 finding is due to factors other than pure chance at the level of .05 and reported that the addition of "Intonation", "Vocabulary", and "Content" adds significantly each time to the R^2. We also understand from the table that the four variables together account for more of the variance in the L2 spoken proficiency test scores than only one on its own.

Such a reading, however, also needs to be seen in the light of the original correlations. Here, we understand that — once error correction is included —

the other variables add little. Thus, error correction comes out as having the dominant relationship. Nevertheless, if we look at the correlations of all the variables with each other, which is always advisable, that outcome might have come about because the other three variables have only relatively small correlations with the dependent variable (or because they have strong correlations with the dependent variable, but *also* strong correlations with error correction). For this reason, once error correction is included in a stepwise multiple regression (and untamed computer programs have a habit of inserting the strongest correlation with the dependent variable first), these three variables get discounted. Thus, we would be looking to the researcher to explain the procedures adopted since, by the nature of such stepwise calculations, the order in which data are entered into the equations will affect the amount of contribution assigned to each (and combined) variables(s). If the researchers chooses to enter the weakest (significant) correlation first (and we should expect this reversal of the norm to be explained), there will be no difference seen in the R^2 total result, in that the four variables here will account for 48% of the variance. However, the reversed order of entry may mean that the cumulative weights differ.

T-tests

Perhaps one of the most used, and abused, of statistical procedures in our field, the t-test compares the means of two (only) groups and determines the confidence or significance level with which the researcher can describe the two samples as different due to some intervention or genuine difference in whatever variable created the groups — rather than pure chance. The first thing to remember here in our appraisal is to establish which of the several t-tests has been applied and whether this is appropriate. In theory, this decision will have been made obvious in the previous section and checked there by the reader as a result of using the flow chart. However, the specific test may not have been named there as such by the researcher, and it would be wise to check identification here. It is reasonable to expect the researcher to specify the t-test applied, both because of the variety available, and because there are subtle differences involved in each as regards the kind of data (and assumptions) that must be applied. Thus, as with all the procedures examined, the kind of t-test used should be viewed in the light of our appraisal of specific assumptions associated with these kinds of tests in the previous section. As in most of the descriptive or inferential statistical procedures reviewed here, an alternative non-parametric

test will be available if a key assumption has not been met. However, the reader would again be advised to check carefully against the flow chart and table to establish the justification for any choice. If the researcher has chosen to ignore a particular assumption and go ahead with a specific test, the justification should be made clear. The appraisal is of vital importance: an incorrect choice of test can easily lead to a serious Type 1 or Type 2 error being committed. In the former, the researcher rejects the null hypothesis when he or she should not have done (e.g. he or she claims the experimental and control group were significantly different when they were not). A Type 2 error is committed when the null hypothesis is wrongly accepted as true (e.g. he or she says the groups were not different when they really were).

Results are typically presented as a statement or in tabular form: "*t = 2.874 (34) (or df = 34), p < .05*" or

Group	**n**	**Mean**	**s.d.**	**t obs**	t crit	*df*	*p*
Experimental	19	15.6	3.7	2.874*	2.042	34	$p < .05$
Control	17	11.4	5.3				

In the above, "Group" denomination should correspond to those being compared according to the previous text (e.g. "*Control*" and "*Experimental*" or "*Normal*" and "*Test*"). Check then to see that the numbers quoted correspond to what was mentioned in the previous "Subjects" section. Any differences should have been explained in the text. The "Mean" and "s.d." will need to be assessed immediately as descriptive statistics in the way mentioned earlier, particularly in the light of what they reveal about the normality of findings. The reader needs to be able to visualise the typical "bell-curve" design of the normal distribution and remember that — in such a distribution — there will ideally need to be room on either side of the mean for at least two s.d. measures. In these fictitious results from a placement test with a total possible score of 28 points, we can use the data immediately to visualise the curve. If 15.6 was the mean for the experimental group (and 28 points was the maximum they could have obtained) and 3.7 was the s.d, this would allow for about three s.d measures on either side of the mean (i.e., $15.6 + 3.7 + 3.7 + 3.7 = 26.7$). We might notice from such estimates that there is actually room for four s.d measures below the mean for the experimental group, which might alert us to the possible presence of outliers in the data. As regards the control group, we note a relatively larger measure of variability, and the fact that these results would also comfortably allow for three s.d. measures beyond the mean but only two below

the mean. Reflected in the normal distribution (and assuming both groups did the same test!), this may indicate that more subjects did rather better than expected in the control group despite not having received the treatment — this might have pulled the mean somewhat higher than it would otherwise have been. If it is felt this may have amounted to a serious threat for a t-test that assumes normality of results, we might hope to see the matter addressed by the researcher. Furthermore, this, (along with any other information drawn from such descriptive measures) would need to be considered carefully when appraising any results based on these data.

We then go on to read a "t obs" and "t crit" figure, and a measure for the degrees of freedom again. This latter adjustment to the sample should correspond (in the t-test) to the number of subjects (in each group) – 1, or 36 – 2 in this case. The "t obs(*erved*)" figure will be that observed after applying the data to the specific t-test formula used. It might be worth checking (if possible) the information given in the relevant statistical table. Once again, in the absence of a row for 34 *df*, the researcher might have applied the more conservative 30 *df* and obtained a "t crit(*ical*)" value for a probability of < 0.05 of 2.042. Since the "t obs" exceeds this value for the level of probability selected, the researcher is right to reject any null hypothesis that said there was no difference between the groups. As we have mentioned on several occasions already, calculating significance of results is a relatively simple matter nowadays with the use of computers; interpreting what surfaces from such calculations, however, requires much more critical insight. The reader will often be presented here with sets of tables and figures that may culminate in the kind of outcome illustrated above and in the other sections of this chapter.

Nevertheless, the hard work of interpreting these figures begins now. For the reader, their own critical expertise will also need to be to the fore at this point. Our reading of any results section will have three objectives. Firstly, there has been the constant need to form our own initial impressions of what the data are telling us, before the researcher gives us their reading. Now, we need to read about what the researcher understands from the same data. Finally, we would want to compare these thoughts with our own reading of the results.

What is your appraisal of any other interpretations that the researcher makes of his or her data in this section?

As we read and react to what the researcher interprets from the results, it will be as well for us to bear in mind what these results can reasonably convey and what they cannot. Firstly, the application of a descriptive or inferential statistical

procedure such as the t-test may tell us whether a significant difference exists or not between the two groups involved, but it does not — in itself — tell us why that difference came about. That is a question of interpretation and one in which we can only have confidence to the extent that the research has been well designed and the relevant threats met. For example, if we are reading a study in which the two groups involved were drawn from intact classes, and no pre-testing had gone on prior to any intervention to establish equality in the groups, we will understand that a statistically significant difference now does not mean that the groups were not already significantly different at the outset. Similarly, we need to understand the nature of the two groups who are now being seen to be at variance. Ask how normal it is to find such a difference between these groups. A study that reports a significant difference between the success of a group of native speakers correcting their output and a group of foreign learners doing the same may not be reporting a difference that is very meaningful. Any difference reported as a result of the t-test may be due to the natives' overall better language ability rather than the specific "correcting-ability" variable. Once again, the focus of attention in our appraisal of both results and the researcher's interpretation of them is the real practical import or meaningfulness of any statistically significant outcomes in the context of the study (see below). Finally, it might be as well to confirm whether any significant results represent the major question(s) being asked or some subsidiary one posed after seeing the results themselves! There can be a temptation to take part in "fishing trips": other, new differences or relationships might be "discovered" in the data obtained that were not previously part of the objectives.

Ironically, non-significant results might — in some cases — be just as interesting as significant ones. We would hope to see the researcher address any outcomes, of course, but we may often find less attention being paid to results that do not come out exactly the way the researcher might have hypothesised. This is a pity, because interesting trends might be evidenced by results that — although they do not quite make the cut-off point required by the proposed (alpha) level of probability — indicate that some effect or relationship of a variable is being observed. Imagine, for example, that a researcher wishes to see whether their revolutionary new method for teaching L2 vocabulary actually succeeds in making a difference to groups of learners. Although the design of the study — using intact classes and with no possibility of random selection or assignment to groups — will not allow for generalisation of any results, the researcher will use the t-test to support the description of any outcomes after applying the method in this context. The researcher opts for an alpha level of

< 0.05 for any calculations, and the t-test shows that the differences between the groups on the post-test of vocabulary are not significant at this probability since the computer reports a significance level of $p < 0.058$. Although the result means that the null hypothesis (that the two groups are not different) cannot be rejected, it would be important not to throw the baby out with the data. The field needs to be informed about the trend here towards a significant effect of the new methodology, for it may mean that the researcher or the reader can suggest ways forward in further research that counteract any possible weaknesses in the design and that could help to achieve "better" results.[4]

This said, we do need to be wary about any results (in this or any other statistical procedure) that are described by the researcher at various levels of probability. The "p" measure essentially tells us about the probability of a particular result and, thereby, whether we can have confidence in rejecting the null hypothesis. Conversely, it tells us nothing about how much of the difference between the two groups is actually due to the effect of the independent variable (see below, "effect size"). This also means that a result that turns out to be at a higher level of probability than the alpha level previously established for the calculation may not be more effective or important than a result that is "only" significant at the predetermined alpha level — and should not be claimed to be so. Similarly, the impression should not be given the reader that a particular result is "very" significant or "less" significant than another. Likewise, the probability level does not tell us about the probability of repeating this result in another context. The best way of determining the replicability of a result is by replication itself!

The bottom line in this kind of hypothesis testing is that the confidence obtained in results by the use of a particular statistical procedure will only give the researcher (and the reader) part of the story. Such "unfinished stories" are also, of course, what drives research forward, in the sense that they show us what other questions need to be followed up in order for us to get a clearer picture of what is going on. However, another aspect of the story behind a

4. It is also worth remembering that there are occasions when a researcher would, indeed, choose to accept the null hypothesis (i.e, rather than not be in a position to reject it). Frick (1995) suggests that too many researchers see their inability to reject a null hypothesis as an indication of "failure". However, for example, there are experimental situations wherein it would be important to confirm that there is, indeed, no difference between two sets of data or groups (Frick, R. 1995. Accepting the null hypothesis. *Memory and Cognition.* 23, 132–138).

statistically significant result is that the outcome will always be influenced by other factors within the design itself. So far, we have looked mainly to validity factors as of fundamental importance to interpretation, but further information in this section of the paper will also help us to assess the true significance of what has been discovered.

What information is made available or can be calculated about effect size of the outcomes?

Therefore, our critical focus now needs to move beyond the significance level obtained in any statistical procedure. Logic should tell us that the independent variable is unlikely to be the sole cause of any significant differences between the groups. As we have already seen in correlation procedures, there will always be some "error" involved in the variable relationship, often due to the inherent nature of sampling. Further research will be needed to determine if any other variable is at work but, for the moment, it would be interesting to see estimated the size of the effect of the independent variable or the **strength of association** between variables. Such a measure would again help us better to appreciate the real consequence of this variable in the relationship or effect studied.

A problem with the kinds of descriptive or inferential statistical procedures we often read about in papers is that outcomes are inevitably influenced by the sample size — making it difficult to assess the true effect or relationship achieved. For example, as can be appreciated from the statistics tables, it becomes "easier" to reject a null hypothesis when more subjects are involved. Ideally, therefore, we would need a measure that gives us an idea of the real effect obtained, without regard to sample size or to the *p* value obtained. A first rough estimate can always made by checking on the difference between the means of the groups. Ask whether that difference seems to be a lot of scale points/marks in relation to the length of the scale: for example, a difference of 1.2 is quite big on a 5 point scale, but not on a percent score scale.

A number of more formal effect size measures exist that correspond to a particular statistical procedure, but these can be translated to the other comparable measure with little difficulty. Sadly, they are rarely used or presented in studies in our field, and it is more common for results only to be taken as far as

the acceptance or rejection of a particular hypothesis.[5] This is unfortunate, since the normal tests of statistical significance will only succeed in giving us part of the story — an indication of the existence or absence of relationship or effect from the independent variable. With information about effect size, for example, we would be able to summarise a series of experiments that used the same independent variable and then directly compare effects across these studies, regardless of the numbers of subjects involved. Similarly, we might average these effect sizes across several studies to provide an estimate of the overall effect; such information would be particularly useful in an applied field like ours, where we so often need to understand the effectiveness of recent innovations and new methodologies and where second language learning goes on in so many different contexts.

In the case of the t-test, and in the absence of information from the researcher, effect size can be calculated by the reader in two ways. Firstly, we can use Cohen's *d* to work out the effect size for a between-groups design simply by $2t$ divided by $\sqrt{df}$.[6] The larger the effect size actually is, the easier it is to detect (i.e., fewer subjects will be needed to detect it). Cohen (1992)[7] (p. 156) describes a medium effect size for a two-group design as being .50, and small and large effects as .20 and .80 respectively. Knowing the amount of effect a particular variable had will help the researcher (and the reader) in other contexts to decide on whether to apply this treatment in other less constrained situations. Information about effect size is also extremely useful to the field when other researchers are trying to establish the most powerful statistical test to use for their study. The researcher will want to have selected a test that is powerful enough to give him or her the assurance that a Type 1 or Type 2 error will not be made in reporting results (see above).

A second class of effect measure can be based on "strength of association" calculations (cf., "strength of relationship" in correlation procedures), which

---

5. It is interesting to note that a growing number of learned journals in the field of applied linguistics, among them *Tesol Quarterly* and *System*, now recommend authors of statistical studies include a suitable measure of effect in their results (see *Tesol Quarterly* , "Information for Contributors").

6. More detailed discussion of effect size can be found in Cohen, J. 1988. *Statistical power analysis for the behavioural sciences.* Hillsdale, NJ: Erlbaum and Rosenthal, R., and Rosnow, R. 1991. *Essentials of Behavioural Research: Methods and data analysis.* New York: McGraw Hill.

7. Cohen, J. 1992. A power primer. *Psychological Bulletin*, *112*, 155–159.

determine how much of the overall variability in the data can be accounted for by the independent variable. For the t-test, we can use ***eta-squared***: $t^2 / t^2 + df$. If we try this out with the results above, we obtain an *eta* of .19, which is not a very strong association. It tells us that 19% of the variability in this study is accounted for by the independent variable. At this point, the researcher and the reader would need to re-consider the objectives of the study. It may well be that the specific statistical test proposed aimed to discover the existence of a significant effect or relationship as a result of applying some or other theory to a set of data. In such cases, even a relatively weak effect would be worth our notice in that it shows us that the theory was confirmed in practice, and such a conclusion should be addressed in the text. At other times, such outcomes will not be those expected as a result of the previous work in the area. Once again, we would be looking to the researcher to address the apparent discrepancy. Such considerations of effect size might also helpfully point the field to other questions still to be answered. Here, for example, a significant effect has indeed been obtained at the alpha level of probability required, yet only 19% of the variability is accounted for by the independent variable. This leaves the question open (i.e., yet to be studied) as to what else could have been involved in the effect (i.e., to account for the 81% of "error"), or whether special characteristics of the sample, the context, or the task might have affected the outcome.

While we have already stated that parametric tests are essentially more powerful than non-parametric tests, the question is still none to clear, and other factors directly related to the sample size and effect size need to be taken into account. A common misconception is that power is increased by *any* increase in the size of the sample used (i.e., the bigger the sample, the more powerful one can claim the study has been). It is true that the more subjects involved, the more likely that the statistical test used might detect a significant effect or relationship, since there will probably be a more representative sample. However, ideally, researchers would have performed a power analysis before collecting data, to estimate the ideal sample size requirements in their study. To do this, the researcher would need to have first estimated the effect size anticipated for the study. The larger the effect size, the greater the power in the study because less extraneous error will have entered the relationship between variables. Here is where the importance of reporting effect size becomes evident, for an examination of the effect sizes obtained from previous studies of the independent variable can guide this estimate more clearly.

What initial conclusions do you come to about the practical significance and meaningfulness of these results? Do these coincide with the researcher's interpretation?

Throughout the text in this section, we will return to the need to appraise results of any descriptive or inferential statistical procedure in a way that separates their statistical significance from their practical significance or meaningfulness. It should already be clear that a so-called "statistically significant" relationship or difference could turn out to be of little import once placed within the wider context of what we now know about the study. Obtaining a significant result does not obviate the need for the researcher (or the reader) to assess that significance on a wider scale. In most cases, such assessment will be a matter of common sense. Remember that statistical significance can still be obtained with a minimum relationship between the variables or with a minimal difference between the means of two groups. Stand back and ask, for example, how we would assess the meaningfulness of a significant difference of a few points between the means of two groups when the test itself had a total score of 150 points? After considering factors such as the nature of the subjects involved, the experience they had had with this kind of test, their level of knowledge, the kind of methodology followed in class, the content of the test and the procedures used to administer it, or even who did the marking and how, we may conclude that this difference does not really mean very much.

Imagine we are reading through the table of t-test results above. We have read previously that the significant differences in means in that table come from a study in which two groups of advanced-level Spanish-as-a-foreign-language students (N = 36) were tested on a multiple-choice type reading comprehension measure (total 28 points). The control group received normal tuition in class, which involved twice-weekly text analyses and multiple-choice questions to test comprehension. The experimental group received the same instruction plus one hour extra for ten weeks during which they wrote their own multiple-choice questions on these texts, received feedback on these from the teacher, and tested them out on each other. Do we feel that the eventual differences shown in the post-test are sufficient to be meaningful in this context? Would we have expected greater differences between them? Do we feel that the numbers involved in the study were not large enough to read much in to what happened? Do we even think that, at this level of L2 proficiency, reading comprehension is not meaningfully determined by multiple-choice questions only and, therefore, little of worth is revealed here about the abilities of the different groups to

comprehend the L2? All such issues are perfectly legitimate reactions to the study from the reader's own reality. Clearly, the researcher will not be in a position to predict or address all of them. We would want to see, however, some attempt to place results in their true context, either in the present section or in the subsequent "Discussion".

Analyses of variance

If a researcher conducts a number of t-tests on the same subjects across a number of independent variables, the chance of making a Type 1 error increases as the number of group means being compared increases. The versatile **ANOVA** can also involve a comparison of group means but has no limitation on the number of group comparisons that can be made (always assuming, of course, that such comparisons are previously planned rather than the result of "fishing" about in the data available!). Researchers use ANOVA to examine the variability of scores within and between groups. Subjects' scores within the same group will vary due to individual differences and random error. If there is a treatment effect found, there will be more variance between the groups than within them.

When a one-way ANOVA is to be used, we would be looking to see that the researcher has proceeded firstly to apply an omnibus "*F* Test", which will result in a **ratio** of the two sources of variance — between-groups divided by **within-groups**. The test will have determined whether any group mean is significantly different from any other group mean. The researcher would be looking for an *F* of considerably more than 1 to show such a difference between the groups and to be able to proceed. If the final ***F* ratio** is not significant, there is no point in going any further, since the test has shown that all the group means are basically the same and the null hypothesis (i.e. no difference between the groups) cannot be rejected.

Results of the *F* test will normally be presented as a statement or in tabular form: "$F = 0.126$ (3, 92) *ns* or $p = 0.945$" or:

Source of variance	**SS**	***df***	**MS**	***F***	***p***
Between groups	6.688	3	2.229	.126	**.945**
Within groups	1620.006	92	17.608		
Total	1626.694	95			

As always, we should first try to understand what the columns and rows are telling us, before assessing what is within the table itself. "Source of variance" informs us about the two sources of variance mentioned above. The total variance represents the sum of the variance between and within the groups. The "SS" and the "MS" columns are the "sum of squares" and the "mean square" respectively, and are steps in the calculation of the *F* statistic. As always, it would be as well to check on figures as far as possible. The MS between groups should have been obtained by dividing the SS between groups by the *df* between groups (i.e., 6.688/3). The MS within groups is obtained by dividing the SS within groups by the *df* within groups (i.e., 1620.006/92). The *F* ratio will have been obtained by dividing the MS between groups by the MS within groups (i.e., 2.229/17.608). There are a number of *df* measures in the ANOVA procedure. The *df* between groups is equal to the number of groups or independent variables minus one (i.e., $4-1$). The *df* for each group is $n-1$. Since in this fictitious study there were 24 subjects in each group, there are 23 *df* within each of the four groups. Because all four groups are the same size, we can obtain the within-groups *df* by multiplying the *df* within each group (23) by the number of groups (4). The obtained *F* value here is less than 1 and the outcome is not significant (*ns*). In this case, therefore, we cannot assume that there is any difference between the groups, and the null hypothesis cannot be rejected. Here the story might safely end, at least as far as the results of analyses are concerned (we would obviously be looking to see even non-significant outcomes addressed in more detail in subsequent sections).

If an *F* ratio of more than 1 is obtained, the researcher must go on to decide if the null hypothesis can be rejected at the alpha level required. We, too, can check this using the **F distribution** (ANOVA) statistical tables and see what level of significance is obtained. In this table, slightly different from that seen so far, the *df* line is horizontal and vertical. The vertical *df* corresponds to the within-group measure and the horizontal line across the top to the between-group measure. When the calculation produces a statistically significant result, the researcher will be able to state with some confidence that manipulating the independent variable has produced a change in performance. This may again be where the story finishes as far as the researcher is concerned. The reader, of course, would now be interested in the interpretation of "significance" discussed above in the "t-test". However, the *F* measure only allows the researcher to reject or fail to reject the null hypothesis; if the result was the rejection of the hypothesis, one of the key questions still remaining is *where* that significant difference is to be located. In other words, the reader

would be looking to see further analysis made of any significant outcome here. The descriptive statistics (hopefully) provided — particularly mean and standard deviation — will be important here for they will provide a first suggestion as to where the differences might be found. However, more precise procedures should then have been used.

The general procedures and logic for designs that use repeated measures are similar to those described here. The main difference is that the F ratio is obtained in a different way to allow for the fact that we are not interested in how subjects differ from each other, but rather in the way each subject performs differently across the different measures or at different times.

Source of variance	SS	*df*	MS	*F*	
Subjects	32.87	5		34.96*	*$p < .01$
Presentation rate	249.62	2	124.81		
Subjects × Presentation rate	35.74	10	3.57		
Total	318.23	17			

These fictitious data show results from six subjects all experiencing three experimental conditions: learning different lists of L2 words under slow, medium, and fast rates of presentation. There were 6 subjects so the *df* is n − 1. There were three modes of presentation so the *df* is n − 1, and the "Subjects × Presentation" simply multiplies the two numbers. The SS outcomes were obtained as a result of previous calculations; we might like to stop and think how the MS figures were obtained from the SS and *df* statistics. Note that the MS for subjects is not required in such a design. We do not expect students to differ from one another or, at least, this is not the objective of the present exercise. Therefore, unlike the between-groups one-way ANOVA, the MS within groups will not be used to determine the F ratio. Here the ratio is calculated by dividing MS between groups (i.e., the presentation rate) by the MS of the Subjects × Presentation. The same distribution table is used to check on the significance of *F*, except that the two *df* to be consulted there are those for the Presentation rate (2) and the interaction (10).

We should expect to find "post-hoc" comparisons made between the group/condition/treatment means. This is because a significant *F* means that at least one group mean is significantly different from another group mean. Two routes are open to the researcher: the first one depends on previous hypotheses being made about differences between groups and the second on the fact that

the researcher began with a null hypothesis of no differences at all between any of the groups. In the first case, there would have been specific null hypotheses made about group differences. For example, perhaps the researcher hypothesised that the Group 1 mean would be significantly different from the Groups 2, 3, and 4 means, or that the means of Groups 1 and 2 would be different from those of Groups 3 and 4. The reader needs to be especially attentive here since the temptation to set out on another "fishing trip" looms at this point: this refers to the highly-unscientific habit of looking around in the data obtained for significant results. In the present case, we would want to see that any subsequent post-hoc tests were carried out based on the previous hypotheses of the study. Thus, the point would need to have been made earlier that, say, Groups 1 and 2 would obtain significantly higher results than the other two groups on a test of L2 speaking as a result of some intervention or treatment. If the researcher opts for the second route mentioned above, the understanding is that, after rejecting the null hypothesis, the researcher is merely interested in pinpointing the precise location of any differences and so comparisons are to be made. There should be no suggestion of tests being undertaken after receiving the information from the descriptive and inferential statistical procedures that there are, indeed, significant differences, which the researcher then proceeds to "find". Hypotheses are — by definition — not made after we see the results, and/or claims then made that this was what was expected from the outset!

A number of these tests exist and each has its advantages and disadvantages for the researcher. The most popular tests are the **Scheffé**, the Tukey, and the Fisher exact tests. With each test, there is a risk run of making a Type 1 error, although this is less likely with the Scheffé test since this is the more "conservative" of the three in that it requires the most rigorous criteria for significance. We should expect to see the particular test named so that the specific consequences can be assessed for the data obtained. We would also expect to see results from these procedures presented in a way that group comparisons can easily be made, ideally in tabular form.

So far, we have been looking at the results from a "One-way **balanced** ANOVA": the researcher has used only one dependent variable and only one independent variable with three or more levels. However, analysis of variance is a particularly popular procedure in our field precisely because it enables the researcher also to study the effects on the dependent variable of a number of different independent variables at the same time. The advantage of such "Factorial designs" is that the researcher will now be able to see both the effect

of each of these variables separately, and also how different variables interact to produce a particular effect on the dependent variable (cf., the procedure of multiple regression, where we can see which one, or combination, of a number of independent variables best *predicts* results on the dependent variable). The simplest factorial design might see two independent variables each of which is manipulated at two levels: a so-called "2×2" design. Crucially, these designs give the researcher the possibility of identifying interactions between independent variables. An interaction will be seen to have occurred when the effect of an independent variable on the dependent variable is determined by the level of another independent variable. In any factorial design, therefore, it will have been possible for the researcher to have made and tested predictions about the overall effect of each independent variable while ignoring the effect of the other independent variable. This overall or main effect will then be interpreted by the researcher. The reader should bear in mind that such interpretation is critically dependent on whether a significant interaction has been revealed or not.

When a researcher "adds" a moderator variable (see p. 24), he or she would be testing whether or not the main effects (perhaps already studied in previous research) of the independent variable on the dependent variable are moderated by some other factor. In theory, he or she might hope cause and effect are clear-cut between the main variables and that nothing else *is* involved. When an interaction is found to be significant, any assessment of the relationship shifts to focus on that interaction, rather than on any significance of the main variables. In general, main effects need to have been interpreted with care whenever a significant interaction is evident. When no interaction has been observed, the main effects of each independent variable can be interpreted as separate relationships, as though they were part of two different studies.

As in all the statistical procedures we will see, the initial focus for appraisal would again be found in basic summary (descriptive) statistics such as the *mean* and *standard deviation* for the groups involved. These statistics are always extremely useful to help the reader come to initial conclusions about what happened:

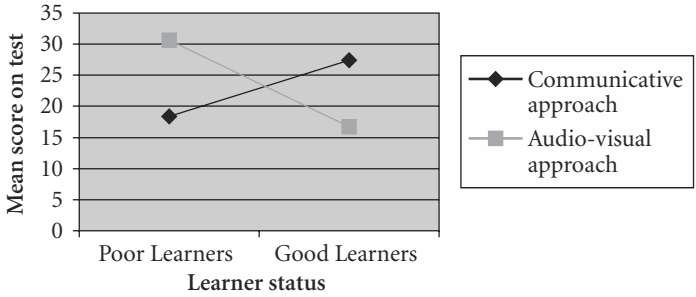
		Learner status	
		Good learners	*Poor learners*
Method	*Communicative approach*	n = 6 Mean = 27.3 *s.d.* = 1.58	n = 6 Mean = 18.5 *s.d.* = 1.87
	Audio-visual approach	n = 6 Mean = 17.4 *s.d.* = 1.14	n = 6 Mean = 31.2 *s.d.* = 1.30

Means for main effect of learner status:
"Good learners" = 27.3 + 17.4/2 = 22.3
"Poor learners" = 18.5 + 31.2/2 = 24.8

Means for main effect of method:
"Communicative approach" = 27.3 + 18.5/2 = 22.9
"Audio-visual approach" = 17.4+31.2/2 = 24.3

By eyeballing these results, we should be thinking that the means in each section of the matrix do indeed seem very different. For example, the poor learners' results with the Communicative approach are some twelve points below those they obtained with the Audio-visual approach. Having said that, our attention should also be drawn to the fact that there are only 6 subjects in each group (statisticians recommend a minimum of 5 observations per section of the matrix). With such small numbers involved, the differences obtained in the means will indeed need to be large if they are subsequently to be found statistically significant. We should also notice that there would appear to be an interaction present: the highest of the two means for the "Good learners" was with the Communicative approach, and the highest for the "Poor learners" was with the Audio-visual approach. Finally, look to the *s.d.* statistics to provide useful information about the variation in each group. Looking across the four groups, we can see that the largest variation was in the *Poor learners/ Communicative approach* group and the smallest was in the *Good learners/ Audio-visual* section. The small difference between these two values would tell us there was little overall variation between the groups. Furthermore, the fact that the largest *s.d.* in any group was below 2 units on the post-test indicates there was not much variation within any of the groups.

Particularly helpful is a graph, similar to that below, in which we would see a illustrative representation of the means of the two groups of subjects (Poor learners vs. Good learners) on post-tests after experiencing two different kinds of language teaching methodology (Communicative vs. Audio-visual approach):



In this kind of graph, non-parallel lines would suggest an interaction and parallel lines suggest no interaction. In this example, the evidence is that, although good learners did considerably better if they were in the group studying with the Communicative approach, poor learners did better on the post-test if they were studying with the Audio-visual approach. However, the graph also confirms that there was an interaction (e.g., good learners did not perform so well when they were using the Audio-visual approach). Such an interaction means that the researcher cannot conclude that one teaching approach is more successful than another; things are not that clear cut as it might have seemed at the outset. In other words, the main effect of the variables now has to be seen in the light of this interaction. In a 2×2 design like this there can be only one interaction, but in a three-factor design each independent variable could interact with each of the other two variables, and they could all interact together. Therefore, the change to a three-factor design means that up to four kinds of interaction are possible. When the interaction is seen (or calculated) to be significant, the researcher should address the issue above that of the main effect. This is because the relationship of the main effects is no longer simple; in fact, the effects of one independent variable are different at various levels of the other independent variable. We should, however, be alert to any abuse of the factorial design's ability to study a number of different independent variables and levels. It is all too easy to design a **factorial ANOVA** with so many independent variables and possible multiple interactions that it becomes impossible to interpret outcomes with any degree of confidence.

The results from a factorial design are presented in a similar fashion to the *F* test above, as the same basic tool is used in one-way analyses and more complex designs. We would firstly expect to see that the omnibus *F* ratio test had been carried out (see above) and that a significant *F* suggested that at least one relationship of mean scores was significantly different. The follow-up analyses are then needed to pinpoint these differences and interpret the initial

omnibus test. In these 2×2 designs with two independent variables, however, there are three potential sources of variation. Each independent variable can produce a main effect, and the two independent variables can combine to produce an interaction effect. Such complexity is best presented clearly in tabular form in the source table rather than within the text itself:

Source of variance	SS	*df*	MS	*F*
Method	27.08	1	27.08	5.39*
Learner status	15.14	1	15.14	3.01
Method × Learner status	103.75	1	103.75	20.67*
Residual Error (Within groups)	82.01	20	5.02	
Total	227.98	23		

* $p < .05$

The information in the ANOVA summary table is initially most usefully read in conjunction with the means reported in the descriptive statistics table. As we shall see below, the statistically significant *F* for the interaction indicates that the pattern of results across the learners for the two methods is different. As we can read in the descriptive statistics table, the means reveal dissimilar results with each method and group. The means for the main effects of "Learner status" and "Method" below the table also reveal certain differences, although not as large. These relatively small differences are reflected in the significance of the *F* tests carried out for main effects. The three *F* tests were computed by dividing the MS value within groups into the MS for "Method", "Learner status", and the interaction. The specific probabilities were not reported as part of a computer program here, and so a table has been consulted (see below) to determine these at an alpha of $< .05$. Thus, the MS Residual Error is used as the denominator of all three *F* tests. The MS for each main effect and for the interaction are used as the numerators for three independent *F* ratios. These ratios then appear in the *F* column. Although we can work downwards from the top of these ratios, it is best to fix our first evaluation on the interaction ratio (if there is one), since this will override any main effects (see above).

Since the F ratios are all more than 1, the researcher can safely move on to the next step to see if this is large enough to allow the null hypothesis to be rejected at the cut-off point chosen. By looking at the table again and *df* numbers (and it is always worth checking both according to the information presented us so far), a value can be obtained for these *df*s for each of three possible effects: method, learner status, and method × learner status. There were

24 subjects in the study and the total *df* is N − 1 or 23. These 24 were divided into four groups, and the *df* within groups is 24 − 4, or 20. There were two levels for "Method" and two levels for "Learner status". In both cases this amounts to 2 − 1, which gives us a *df* here of 1 for each. The *df* for the interaction between Method and Learner status is 1 × 1. The "Residual Error" is the variance associated with normal variability in performance (i.e., all the variance not accounted for by any main effects or interactions). It follows that such variance is not influenced by either independent variable or any combination of the two. In the statistical table, the position where the (between-groups) *df* meets the (within-groups) value (1, 20) gives us a critical measure of 4.35 to reject the null hypothesis at $p < .05$. Since the *F* ratio for "Method" is higher, there is a significant difference obtained at this level. On the other hand, the value obtained for "Learner status" is less than the critical value and not significant. The interesting point to note in this table, however, comes with the interaction value, which is also greater than *F* critical. This should immediately attract our attention, since it means there *is* a significant interaction between the two, as predicted by our initial eyeballing of the descriptive statistics. In other words, any effect generated by the variable "Method" will be moderated by the variable "Learner status", exactly as the graph also predicted it would.

In a complex design, as in the one-way version, some kind of follow-up analysis of this omnibus test is now required. The analysis (and our appraisal) would differ depending upon whether or not a statistically significant interaction has been obtained. If the interaction effect had *not* been significant here, the researcher would be able to go on to address more confidently the main effect of the independent variable. If the main effects of the variables are statistically significant, the source of that main effect can be specified more precisely by performing analytical post-hoc comparisons that compare two means at a time. If an interaction is not obtained in such a factorial design (and always assuming the researcher has conducted a sufficiently powerful and sensitive study to detect it), he or she would have the evidence to be able to support the generalisation of the effects of each independent variable across the levels of the other independent variable(s) within the experiment. Generalisation beyond this would depend on the usual kind of assumptions I have mentioned earlier in the book.

In the present case, we would want to see any significant interaction appropriately addressed and interpreted in this or the subsequent "Discussion and Conclusions" section (see below). The basic reading of the results is that differences revealed in the two teaching approaches have to be consequent

upon the fact that poor learners performed better if they were exposed to the Audio-visual approach and good learners did best if they were exposed to the Communicative approach. Such findings might suggest that (in this context) there is an argument to have two methods for these two streams; the practicality of that would need to be assessed in the paper. At this point, if the researcher wanted to identify where precisely any differences were in the groups, we might again be looking for post-hoc comparisons to be made between means, since there is at least one significant difference somewhere among the group means for "Method".

We would be looking to see the researcher address the nature of any relevant independent variable (i.e., one that has been seen to affect the dependent variable directly (in a main effect) or indirectly (in an interaction with another variable)). However, we may also read of the implicit dismissal of an independent variable as irrelevant simply because no effects were revealed. This would be another example of accepting an unfinished story. The advantage (or disadvantage!) of using complex factorial designs such as these, where a number of interactions might take place, is that the effects of a particular variable might well be hidden from view in other — as yet unstudied — interactions. It follows that we should think carefully before rejecting the effects of such a variable; a number of explanations for what happened still remain for further research to study, and it would be useful if the researcher points these out to the field. For example, it may be that a particular independent variable might well have had an effect if *other* levels of the independent variable had been tested. Furthermore, if an independent variable has no apparent effect in a 2×2 design, this is not to say that interaction will also be absent in a more complex set-up. Finally, there is the problem we have continually mentioned in these analyses: failure to reject the null hypothesis (or the absence of a statistically-significant result) does not necessarily imply the absence of any effect. It may well be that the researcher needs to check on the power used in the study and that a readjustment of this could bring about differences in outcomes in the future.

Once again, therefore, we would benefit from seeing effect size or strength of association calculated in some form or another in an ANOVA study. There are two basic choices for effect size, depending on whether the analysis is between-groups and/or has equal *n* sizes, or is repeated-measures and/or has unequal *n* sizes. The **omega²** statistic is best calculated for the former and eta² for the latter case. I will again include their calculation here, for effect size is not often referred

to in our studies, and it may again be up to the reader to measure this.[8] There are three possible calculations here for ω^2 : there were two main effect variables and one interaction present here. However, as a result of our calculations above, only the effects for "Method" and the interaction were significant; therefore, we do not need to calculate effect size for "Learner status".

$$\text{As } \omega^2 = \frac{SS_{\text{method or interaction}} - (df_{\text{method or interaction}})\ MSW}{SST + MSW}$$

For "Method" = 27.08 – (1) 5.02 / 227.98 + 5.02 ⇒ 22.06 / 233 = **.09**
For the interaction = 103.75 – (1) 5.02 / 227.98 + 5.02 ⇒ 98.73 / 233 = **.42**

It is now evident that the effect size for "Method" only accounted for about 9% of the overall variance, and this despite its significance in the ANOVA analysis. On the other hand, there was a much larger effect size for the interaction of about 42%. Therefore, the interaction appears to have been far more crucial than simply the method on its own. Having said all this, and noting that the effect size of "Learner status" would have been very small indeed, the fact also remains that nearly 50% of variance remains unaccounted for in the design. We would want to see this addressed somewhere in the text — perhaps with some suggestions from the researcher about how future research might follow up such an outcome.

We may also come across a number of more advanced uses of **multivariate analyses**, in which the design allows for the study of more than one dependent variable, which are thought to be related to one another. MANOVA is able to examine, say, two or more groups on one or two tests. As with t-tests, it is inadvisable to opt for a separate ANOVA on each dependent variable, as if these were unrelated. To do so would easily lead the researcher into committing a Type 1 error. Using MANOVA has the additional advantage of allowing the researcher to examine the relationships among the dependent variables and determine how the independent variable relates differentially to each dependent variable. Although space does not allow us to describe these more advanced techniques in detail, it is important that the reader understand the

8. eta² for un**balanced designs** and repeated-measures designs is easily calculated from information in the omnibus F test = $SS_{\text{between-groups OR factor of interest}}$ / $SS_{\text{total OR effect + interaction}}$. If the table is not presented, it can also be obtained from the F ratio for the between-groups effect: $F\ (df_{\text{effect}})$ / $F\ (df_{\text{effect}}) + df_{\text{error}}$. Either way, omega² tends to give a more conservative estimate of effect than eta².

general similarities between all these kinds of multivariate tests. Once again, two steps should have been recorded. An initial omnibus test will have checked to see whether there are overall differences between groups on the combined dependent variables. If the answer is affirmative, and the difference is significant, the second step will be to see where the differences are, using follow-up tests. The reader is referred to the handbooks in the reference list for the procedures, advantages, and disadvantages of each follow-up test, the most common being univariate ANOVAS or discriminant analysis. Similarly, MANCOVA might be used to study the effect of an independent variable(s) on multiple dependent variables while controlling for other variables that are predicted to be related to the dependent variables. The principal aim of such analyses is to reduce the error by controlling for the relationship between the **covariate** and dependent variable.

Chi-squared

Chi-squared is one of the few statistical procedures that allows the researcher to address relationships between two nominally-measured variables. The analysis basically compares the frequencies obtained with other **expected frequencies** (i.e., if there had been no relationship between the variables) and sees how much variation from the predicted distribution is normal, and how different these should be for a conclusion to be reached that there *is* a relationship between the variables tested. As with all the analyses seen so far, the researcher should have met the critical assumptions associated with the specific procedure before analysis was undertaken. In this case, the two highlighted are independence of groups and of observations. However, the reader might also want to consider other elements of the test, which — while they are not strictly speaking assumptions — might affect the reading of findings here. Firstly, Chi-squared will at best only reveal significant differences in frequency data and thereby test the relationship between the variables. There can be no suggestion made that Chi-squared in itself reveals a cause-effect relationship, however. Also, although the implicit assumption that frequency/categorical data be used seems clear, it is worth checking to see how far these data are a fair measure from the whole sample. For example, we would want to be sure that frequencies were not artificially inflated because some subjects were somehow able to contribute more to the results than others. In a study that looked at the relationship between introversion/extroversion and participation in class, for example, our

experience might tell us extrovert subjects would normally participate more often than introvert subjects. Thus, any comparisons of the frequency of participation might well present a specious relationship. In such a situation there would be a major threat to the assumption of independence of observations, since subjects would have been able to contribute more than others in each frequency cell. Chi-squared is also a popular procedure to study survey data, which generally involves frequency data that are inappropriate for the other types of data analysis illustrated above. We would want to know how representative a frequency count of, say, the number of people answering "yes" or "no" to a certain question really was. If the question had been asked of 500 students in a telephone survey, for example, we would want to know what percentage actually responded to (rather than refused to answer) the question, in order to know whether the response frequency reflected fairly the number of people who were actually asked the question.

Secondly, it is worth checking to make sure that observations fall into mutually exclusive and/or logical categories (e.g., male/female, left-handed/right-handed, etc.,) rather than counting towards more than one frequency tabulation. We might also come across suspicious data-transformation, where categories have been set up based on data so specific to one context (such as converting scores on a reading comprehension test to "high", "low", and "intermediate" data) that they can no longer be generalised to other contexts. In other words, the extent to which the categorisation is logical or acceptable will need to be related to the research question or hypothesis itself. In turn, we will be looking to the researcher to justify this and also consider its consequences for any eventual generalisation of outcomes.

Finally, we would want to be sure that each expected cell frequency had a minimum of five. If this is not so, and/or if the study carries a *df* of 1, the researcher should be seen to have applied a correction factor (Yates') to compensate for the discrepancy that arises in these circumstances between the Chi-squared distribution and the observed Chi-squared value.

In one-way chi-squared analyses, the researcher is interested to see if frequencies from the dependent variable are significantly different with levels of another independent variable. My focus here will be on the two-way design as by far the more common procedure in our field.

Results from the Chi-squared (χ^2) test will normally be presented as a statement, possibly accompanied by a contingency table and summary statistics: "$\chi^2 = 14.69$, $df = 4$, $p < .05$":

	L2 chosen					
	French	**German**	**Russian**	**Hebrew**	**Italian**	**Totals**
Female	10.08	10.92	4.2	7.56	9.24	42
Male	13.92	15.08	5.8	10.44	12.76	58
Totals	24	26	10	18	22	**100**

$\chi^2 = 14.69$, $df = 4$, $p < .05$.

In this example, the fictitious data come from a study where the objective was to relate choice made by L1 English first-year undergraduates of optional second language in University P to the sex of these subjects. Although chi-squared does not assume any directionality in the variables, we might want to think that there is one dependent variable here (Sex, two levels) measured as a frequency and one independent variable (L2 chosen, five levels) measured as a nominal (**category**) variable. The data refer to the expected values, rather than what was actually observed; in other words, how the frequencies would have turned out if there were no relationship between L2 choice and sex. We would have hoped to see the **observed frequencies** too, as we would then be able to see where differences were greatest and which cell(s) contributed most to any significant outcome. χ^2 uses observed and expected frequencies to test a null hypothesis (here) that there is no relationship between sex and choice of L2 option. The *expected* frequencies of males and females choosing these options are those that would make these proportions the same and force us to "accept" the null hypothesis.[9] To the extent that the observed (i.e., real) frequencies differ from the expected frequencies, the results provide evidence for rejecting the null hypothesis. Once again, we should get into the habit of checking the data presented as far as possible (particularly since the original descriptive data are not provided). Firstly, the "Total" numbers around the table (known as the "marginals") refer not only to the row and/or column totals of expected frequencies here, but also to observed total numbers in the rows and columns in the original frequency table. The *df* here cannot just be calculated by subtracting 1 from the total number of groups. Here the researcher would also have had to subtract 1 from the subject options, of which there were 5. Hence the *df* for the groups is $2 - 1 = 1$ and that of L2 options was $5 - 1 = 4$. The multiplication

9. In theory, we will never have to *accept* a null hypothesis. The outcomes of hypothesis-testing are that the researcher either "rejects" or "fails to reject" that hypothesis.

of these two results gives us the total *df*, which is 4. The observed χ^2 statistic (14.69) reflects the overall size of the differences computed between the observed and the expected frequencies. The greater this difference, the more likely this outcome is to be significant. We are now ready to check up the apparently significant difference in the χ^2 statistical table. At the intersection of 4 *df* and the projected alpha of $<.05$, we see a figure of 9.49. $\chi^2_{observed}$ would need to be equal to or larger than this $\chi^2_{critical}$ (which it is) to be significant at this level of probability. An alternative, if a computer program were used, would have been for the researcher to tell us that statistical probability had been achieved, quoting the actual probability (here, by the way, .0054).

So far, so good. The reader will again need to be alert to how this outcome is interpreted in the light both of the limits of this kind of analysis and of the constraints placed on outcomes by the specific design itself. Our first concerns should be for any possible threats in terms of what was said above. We might first note here that one of the expected cell frequencies was predicted to be below 5 (Female/Russian). The present example is border-line in this respect and it would probably be acceptable to include it in the analysis. We might be looking to the researcher to add a caveat about this low frequency and/or attempt to explain it. Secondly, we need to read carefully the researcher's immediate reactions to these outcomes. Firstly, we would accept the researcher's claim that there *is* a relationship between sex and L2 option. Secondly, we would not accept that being male or female actually causes the choice or vice versa! Other variables (i.e., yet to be studied) might also be moderating the relationship. Thirdly, we would accept this outcome as confirming the description of these subjects in this study, but we would not accept any generalisation to other contexts unless threats to external validity had been seen to be met earlier.

Perhaps most importantly, the researcher is now in a similar situation to that revealed after the omnibus *F* ratio test in ANOVA. He or she knows that a significant difference exists between the frequencies expected and those actually obtained, but does not yet know for sure where that difference was actually located. It might have been in the frequency of any of the cells or a combination of these. Of course, the researcher might not want to go down this road and is content with describing a significant difference somewhere in the data or with highlighting which cells have the biggest expected-observed differences. There is nothing inherently wrong with this; however, he or she would have to understand that a concerned reader might well have reason to be interested in more precise location of differences. In this case, we might be looking out for

"post-hoc" comparisons, whereby each pair or cell is tested for its significance in individual χ^2 analyses. If this information is provided (usually presented in a similar table), and *if* the original (i.e., obtained) frequencies had also been given, the reader would be able to check to see that the outcome is roughly as expected after a perusal of the obtained data. Any apparent anomalies should be addressed by the researcher.

Despite the temptation to finish the story with a "significant ending", it should remain of interest to both researcher and interested reader to discover just how significant "significant" really is in this context! We would once again, therefore, be looking for some calculation of effect size or strength of association to be made, and (either here or later in the "Discussion") reference to the practical significance or meaningfulness of what was found. The two — most common — options are Phi (Φ), when 1 *df* is involved and Cramer's V when more than 1 *df* is computed. Naturally enough, it only makes sense to calculate effect when significance had already been obtained in the previous χ^2 analysis.

As before, the appraisal of the practical significance or meaningfulness of any statistically significant outcome here depends on common sense and recall of what we have been told about the research design and context so far. During our reading, it is as well to keep in the back of our minds that the original descriptive data may actually hold far more informative or revealing data for consideration. In the above example, the researcher might want to compare the χ^2 statistic obtained at this probability level with those obtained in similar studies and go on to address any differences found both in the end result and in the observed frequencies themselves. Similarly, we ourselves might want to see the observed frequencies so that, for example, we were in a position to judge whether any statistically significant difference between, say, the 10.44 males expected to opt for Hebrew as a L2, and the 15 who actually did, represents an interesting result over a sample of 100 students. If we read that the significant contribution to the overall Chi-squared was found to come from the "Hebrew" cell, how might we then react to reading that only 1 female and 17 males chose Hebrew? Equally, of course, any comparisons that turned out *not* to be significant or that came close to the cut-off point might be interesting to ponder in the circumstances of the study.

4. Discussion and conclusions

The quality of the discussion and conclusions

This part of the paper highlights the cyclical nature of any research study: what follows now should be an attempt on the part of the researcher to take the reader back to the introductory sections of the report and, in a descriptive and interpretative summary, show the extent to which findings have answered the questions or hypotheses proposed at that juncture. To that end, the reader him or herself will also need to recall what was said (and appraised) at that point and decide whether what was argued then has now found an adequate and appropriate response in the findings and the subsequent discussion and conclusion. We would also be hoping to see the researcher move beyond the confines of the present research study and show us how the outcomes fit in to a wider context and contribute to the advancement of knowledge in the field. With this in mind, we would be interested to read how the researcher sees their results relating to current knowledge, how this knowledge might have been improved or modified by what has been learnt, whether these results might have practical implications for second language learning, and whether further research is suggested or recommended as a consequence of what has been revealed. Moreover, by laying out the implications of the study for current theory and further investigation, the researcher is simultaneously providing the reader with the kind of information that helps him or her advance in their own study of the topic. Bear in mind that one of our initial objectives in reading the paper in question may well have been to acquire more background data about a specific aspect of the field, prior to mounting our own study and venturing empirically-based interpretations of a particular, related language-learning event.

It should already be clear from this introduction that the "Discussion and Conclusions" section of the paper is a particularly appropriate juncture for both researcher and reader to be engaged in a tacit exchange of opinion about what has happened, what conclusions can be drawn from this, and what remains to

be done. Therefore, we will want continually to be alert to the theoretical and empirical support for any opinions offered here, as well as to the logic of the manner in which they are expressed. Our own thoughts about the research as a whole will need constantly to be compared with those of the researcher. Such reaction on our part is again best realised through frequent questioning of the text: do we agree with the researcher on the general conclusions to be drawn here? Does a particular claim follow logically from the evidence or support provided? Do we think some results support what has been hypothesised earlier, but others do not? Do we feel that current knowledge in the field — as described in the introductory sections — has been appreciably enhanced by what has been discovered? Do we agree with the researcher about what remains to be done and/or how this should be done? Most of these reactions will implicitly reflect the confidence we have gained in the researcher's work as a result of reading the paper. Apart from the research design and execution itself, this confidence will also be based on the logic and strength of argument with which the researcher seeks to defend the outcomes obtained and the conclusions drawn from them.

What conclusions were drawn from the study, and how do these reflect on the original questions and/or hypotheses?

Because this section is basically interpretative in nature, there is no established sequence as such in the narrative. However, ideally we would be looking to read and appraise a number of different perspectives on the findings. The first of these would be a concluding summary of findings: the reader would want to be clearly informed of what the researcher has learned from the study. This statement would ideally include the responses to the research questions, or the support or non-support of the hypotheses as formed in the introductory section, perhaps in the form of a broad statement rather than a mere repetition of the specific results. As we read this statement, we should also be concerned to see that it is consistent with what we have just read in the "Results" section. A researcher may include "new" findings in this conclusion and/or then proceed to over-interpret what the global outcomes mean. Likewise, at some point here, we should expect the researcher appropriately to address any results that have not come out in the way hypothesised or expected. Some kind of explanation might be ventured for these apparent discrepancies, although we should consider such *post-hoc* explanations only tentative, since they would not have been submitted to the empirical testing involved in the main part of the study. Nevertheless, such suggestions are useful for they could form part of any future research in the area

that might seek to clarify the situation more (see below). Finally, once the "Discussion and Conclusion" has been digested, we ourselves might want to think back to the reasons for the study as set out in the "Background to the problem and the problem statement" and consider the extent to which findings are now seen to respond to the problems expressed there.

> **What is your appraisal of the general inferences which the researcher draws from the findings? How do these compare with your own reactions to what you have been told throughout the paper?**

A central focus in this section will be on what the researcher infers from their findings. Indeed, the *APA Publication Manual* encourages researchers at this point to feel free "to examine, interpret, and qualify the results, as well as draw inferences from them." (p. 18). By definition, interpretation will obviously need to go beyond the mere description provided in the previous section. While any interpretations ventured there may have been at a level quite close to the data, now the inferences may involve considerably more abstraction, perhaps to the level of a larger theory (see below). We will now want carefully to compare our own cumulative appraisal of what has come out of the research with that of the author's, in particular regarding what happened within the set-up and conduct of the study to account for the findings. It is crucial for the reader appraising such interpretations to learn how to pause in their reading to consider the reasonableness and acceptability of these interpretations given our own reading of the study. In other words, we will need constantly to ask ourselves whether the explanation or interpretation being offered is reasonable in the circumstances of the study and, if necessary, once in the wider perspective. For example, our earlier appraisal of the statistical procedures used in a paper may implicitly have endorsed the results obtained as a consequence of using that procedure. Nevertheless, the way the researcher then goes about *using* these outcomes in their own interpretations of data will still need to be examined. While we may agree with the results themselves, it is perfectly legitimate to disagree about what they mean. No findings speak for themselves; the researcher will need to interpret these for us and, almost inevitably, will do so from their own particular perspective. The angle from which he or she reads into these data will, by definition, not be the only one available.

A researcher may give us to understand that his or her results are unlikely to have been affected by any intervening variable or other more observable feature, but we ourselves might have identified certain threats on this score in an earlier appraisal. We might then wish to challenge that interpretation of

outcomes on the grounds of its acceptability. Similarly, reasoned speculation is to be expected at this point but, as a result of our reading, we might find it *un*reasonable to see it suggested that results from a small group of intact classes in one school can be appropriately generalised to other schools in that country, let alone further afield. Our interaction with the researcher and his or her interpretation of the evidence should also mean that we are actively engaged in looking for alternative explanations: we might read of the improvement noted between pre- and post-test measures of L2 proficiency but feel that this may be just as easily explained by parallel language-learning experiences as by any special treatment the subjects had been receiving.

In what ways are the findings related to current theoretical and empirical knowledge on the topic?

Finally, the interpretation of any findings could now move away from the confines of the immediate research context to relate the results and interpretation of these to the existing theoretical and empirical literature, much of which will already have been referred to in the previous review. The distinctive cyclical development of a research study and the cumulative nature of research itself can signal a need now to re-assess this literature taking into consideration the present findings, particularly literature cited as part of the background justification for the study in question. For example, if we were told then that gaps existed in current knowledge, we should be interested to know how far these have now been filled by these data. Has consistent or conflicting information been added? Any apparent conflicts or inconsistencies noted with this research should not be left without explanation, of course, and we will again want to consider the adequacy of these clarifications. Indeed, if the researcher does not re-address this work, we might usefully consider what was described in the review of the literature and attempt ourselves to assess the contribution of the present findings. Any pertinent knowledge we might possess of the current state of the field might also alert us to any intrinsic bias here. There may be a temptation for the researcher to seek to defend a particular position suggested by their findings by only citing the literature that supports such a view. If, on the other hand, we have been told of, or are aware of, opposing viewpoints or conflicting evidence, such information could be exploited to enable us better to place results in their true perspective.

In many cases, such considerations of past and present findings might also help the researcher and reader to broaden the interpretation of the immediate findings to place them in the context of any larger theoretical problems ad-

dressed by the project. In this way, we will be better able to assess the contribution of the work to current thinking on the subject and/or link the results to other areas that are theoretically or empirically related to the study. Such attempts to extract meaning and principles from findings are especially useful to the field in studies that have found their starting point or rationale in a particularly theory and/or have set out to provide contributory data to confirm or refute this. We may remember that one aspect of our earlier appraisal of the "Background to the problem and the problem statement" was precisely the likely contribution of the study to current theory. At that point we emphasised — amongst other things — the need for findings to have the potential to help us "better evaluate a number of previous explanations or models" (p. 8). Then, we were looking for references to prior thought or previous theories that directed us to where the present research would fit in with what is already known and, potentially, contribute with further evidence. Now, it might be useful to see this contribution made evident by the researcher, integrating findings into an already-existing theory or model or using them to formulate an original theory or model. Indeed, one of the advantages for the interested reader of a researcher integrating their findings into existing theoretical knowledge is that theories tend to be heuristic: they serve to generate ideas and, by integrating outcomes within existing models or using expected or unexpected outcomes to form new suppositions and hypotheses, theory can stimulate the future research needed to test them.

> **What limitations or weaknesses have you or the researcher identified, and how might any future research seek to contribute further to what has been revealed in the study?**

As we mentioned above, as readers, we will also implicitly have been noting any shortcomings and weaknesses in the study during our reading and appraisal. Now is the time to recall any of the outstanding concerns (particularly those pertaining to reliability and internal/external validity threats) and compare them with the researcher's own views of any limitations on the results obtained or deficiencies in design. The *APA Publication Manual* recommends researchers include in this section "[remarks]… on certain shortcomings of the study, but not [to] dwell…on every flaw. Negative results should be accepted as such without an undue attempt to explain them away." (p. 19). Such statements by the researcher are not to be seen as a signal that the study is flawed, nor their absence interpreted as an indication that the researcher thinks it is perfect! Quite the reverse, in fact, for they demonstrate a researcher who is able to stand

back from the work and, with hindsight, recognise where things could have been improved. It can be safely assumed that no empirical research is perfect. As interested readers and/or prospective researchers, we have much to learn by seeing these shortcomings fully discussed and explained (rather than "explained away", as the *Manual* puts it!) and then followed by descriptive proposals (to which we will need to respond) for additional research which perhaps seeks to correct these deficiencies, clarify any ambiguous results, or test any new hypotheses that are suggested by these findings.

Aside from this perceived need for further research consequent upon inherent shortcomings in the study, we would also be interested to read suggestions that point the way ahead for any further study and research in the particular field. By considering what could have improved the design of the research, by looking for alternative explanations to outcomes, and by comparing evidence from previous studies or existing theories, the researcher may now be in a better position to tell us "where to go from here". This is, potentially, one of the most productive sections of the paper for the reader: he or she can use these suggestions (together with their own ideas) as a guide towards acquiring further knowledge about the topic and, perhaps, as a stimulus for their own research project. Clearly, however, any recommendations we read or suggest should be seen to have developed logically from the findings obtained in the present study. The discussion itself would hopefully have motivated the reader and the researcher to consider questions that remain unanswered along with the kinds of research that would help provide responses to them. When appraising the author's own suggestions, we might hope for something more precise than a sweeping statement such as "future research should try this out with more mature non-native students and over a longer period of time". The reader and the field are much better served if we are given some guidance by the person making the recommendation about *how* he or she hypothesises any future outcomes might vary with these kinds of subjects and *why* more engagement with this population should prove fruitful.

Often, as a result of our reading, the data and the subsequent discussion fail to convince us that there is adequate support for the researcher's position at the end of the paper. In the absence of specific recommendations from the researcher, we might ourselves — as part of our appraisal — then consider replication of the study, perhaps with different subjects, a modified data collection procedure, or design — all with the aim of obtaining more evidence for or against the researcher's conclusions. Conversely, the way ahead may be illuminated by further study of certain aspects of a particular L2 learning

phenomenon brought to light as a result of our reading of the present research. For example, as a result of statistical analyses of data, a researcher might have shown that subjects who have been taught L2 pronunciation in language-laboratory classes acquire better pronunciation skills than those taught with conventional classroom methods. However, as we ourselves read these outcomes, we might have made a mental note about the need for more study of the way subjects actually acquire pronunciation in the language laboratory. In other words, we are not questioning the outcomes here and/or suggesting replication; our appraisal of what happened has sent us thinking about discovering how these subjects might have been processing what they heard to account for their subsequent improvement.

What is your appraisal of any practical inferences which the researcher draws from the study in terms of pedagogical implications or recommendations?

Second language learning is essentially an applied field of research and, as such, studies conducted in this area could helpfully generate some recommendations for modification in educational practice. It should go without saying, of course, that any such proposals for the application of findings must be seen to proceed logically from the actual results obtained or from the general theory or model to which these results have now been applied. Once again, there may be the temptation to seek to recommend the application of findings to other subjects or other language-learning settings without having minimally prepared for such generalisation in the research design itself. Therefore, we would want to think about any such pedagogical implications carefully. This is particularly the case when a "new" methodology, test, textbook, or other learning approach has been the subject of experimental study in the research. In the light of what seem to be significant results from some innovative intervention in the language learning process it is, indeed, appealing to wish to tell everybody the good news and recommend we all take up the new approach as soon as possible.

Nevertheless, much of the research we read will have been carried out in relatively small-scale language-learning operations with few, if any, guarantees of adequate external validity. In such cases, enthusiastic pedagogical recommendations will need to be tempered with the knowledge that there are many, very diverse, second language-learning contexts throughout the world. It would be presumptuous on the part of the researcher to think that his or her intervention will bring about the same success in any situation, let alone to make recommendations about how each such context might best be modified by

taking these findings into account. Indeed, many such specific pedagogical implications are often best levelled nearer "home", the researcher directing the more extensive recommendations of the research towards what still remains to be discovered about a particular phenomenon in the present context before any firm proposals for general practice are made. That the reader is not cognisant of the specific language-learning context to which the results are to be applied should not be an obstacle to our appraising (or making) these recommendations on the basis of their common sense, perceived usefulness, or workability.

For example, we might consider carefully a proposal for introducing to beginner-level students a previously "successful" training course that taught dictionary-use strategies to L2 intermediate-level subjects. Depending on a variety of other factors, we may feel the way an L2 beginner needs to go about using the dictionary (and the type of dictionary itself) may be rather different from that of an intermediate student, and the training might need to be suitably modified before wider application in the context. Likewise, the fact that a study has revealed significant evidence that shows subjects with home access to the Internet enjoying more success as L2 learners may not, in itself, be sufficient even to recommend the immediate massive acquisition and incorporation of on-line computers in language learning classes at a particular school. It may well be that some of the projected recipients of the scheme will need non-specific computer training prior to their using the machines for any explicit language-learning use. Any hypothesised improvements in L2 learning as a result of using the computers may be dependent on how successful this initial training turns out to be. Finally, the researcher should be seen to limit his or her proposals to those that follow appropriately from the present research context. We might want to question, for example, an unqualified suggestion that a particularly successful children's beginners L2 learning programme might be equally successful with young adult beginners.

Earlier, I emphasized the fact that the tacit exchange of opinion between the researcher and the reader — which has, in fact, been the mainstay of our appraisal method throughout the paper — should be even further stimulated by this discussion of results. However, our understanding and appreciation of the appropriateness and logic behind the researcher's conclusions inevitably depends on the way those opinions are expressed. Our appraisal of a paper has assumed throughout the need to address the precision with which arguments, facts, and findings are communicated; however, arguably, here more than anywhere else in the paper, we must pay particular attention to the language used to present an argument or conclusion. It is the reader's responsibility to

think about the argument being presented and then evaluate it. This whole process is less straightforward than it may sound, since it requires us to look rather more closely than we might normally do at the words and expressions used to make claims, present evidence for those claims, and draw conclusions. We will need to be able to identify the strengths and weaknesses of an argument in order to appraise its significance. We will need to establish which elements of the argument proposed are useful and which might need to be discarded or re-phrased. By so doing, we are performing a useful service both to the author and the field and will perhaps be able to re-formulate what has been proposed to produce a new slant on the topic and, thereby, point the direction to new research in the area. Nevertheless, our concern should always be to focus on the kernel of what is being proposed rather than the person making it; the convention in appraisal of another person's work and opinion is that this is treated with due respect at all times.

However, it is worth making the effort here because we will then find it easier to appraise insight, strength, inadequacy, lack of plausibility, or even fallacy in the arguments or conclusions being presented. If we do decide that something has been found wanting in an argument, this should then be explained in a way which makes it clear *what* we have found wrong. There follow a number of typical textual elements of this academic style of writing which may help in an appraisal:

- There is a natural tendency to "hedge one's bets" when presenting conclusions or implications based on empirical data. We might read that something will "*probably*" happen, that "*some*" language learners will benefit from a particular methodology, or that "*generally*" L2 learners will demonstrate this or another problem. These are arguments presented with "qualifiers", words that serve to limit the scope of a claim in order to make it more immediately acceptable and delimit its application. There is nothing inherently wrong with this, and the research design used may actually recommend such a stance, but it does mean that we will have to assess the implications of such constraints for the strength of the conclusions being drawn.

- Look out in the text for words that are used to structure an argument and adopt a posture. It is essential to have the structure of an argument clear in our minds before it is appraised. Basically, we should be looking to ponder both the conclusion drawn and the premises behind it: what needs to believed, or what evidence would we need to have to justify our accepting the conclusion? To be able to do this, we need to decide what can reasonably be admitted as evidence.

These standards are not universal, but rather subject specific in the present case. Our appraisal of what we have read so far in the paper will be our principle guide in deciding upon the strength of argument here.

Words such as "*thus*", "*therefore*", "*hence*", or "*consequently*" can be used to link evidence with claims and suggest inference, reason, and conclusions. Isolate the sentences in which these occur and consider how far the conclusion actually follows from the premise. Taking the conclusion expressed by these words, stop and ask what reasons are presented in the text for believing this conclusion, or why we are being asked to accept the conclusion. Typically, look out for words like "*because*", "*since…*", "*it follows…*", and so on as introductions to reasons. By being suitably sceptical at this point, we will be in a better position to reveal mistaken assumptions, faults in reasoning, and misleading notions in arguments, all of which will help us to build up an appropriate response to what we are reading.

- Look at the claim being made and try to appraise it from a number of angles. For example, call attention to any vagueness in what is being claimed, often observed as a result of referring to something without clearly defining it, or defining something in one way early in the paper and in another way now. Similarly, consider the response to any apparent attempts to convince the reader of the reasonableness of an argument by exaggeration or over-statement. For example, we might follow a particular claim throughout the paper and see whether or not it differs each time it is presented, particularly when the same evidence is being used to support it.

- We should consider the consequences for any conclusions drawn of other typical inaccuracies when constructing arguments. Sometimes, authors over-generalise in their use of language: they may use "*all*" language learners when they mean "some", or "*most*" L2 students, when they mean "those subjects I have studied". This kind of careless over-generalisation has the effect of implicitly discarding or underestimating any contradictory examples — which the reader may then identify. Conversely, we might often need to highlight arguments based on restricted instances of a particular phenomenon. In this case, the researcher may be building up a shaky proposal, as it is founded on unusual or unrepresentative examples. Care also needs to be taken when appraising conclusions presented with the appeal to a respected authority. An author should indeed tell us if their results mirrored those reported by a well-known specialist (or any other researcher, for that matter). It is quite another thing to suggest that the conclusion or proposal gains in strength *because* some

respected author has reached the same conclusion or suggested a similar stance. An argument should be able to stand up on its own because of the evidence produced, rather than because any number of colleagues have implicitly backed this up. Finally, we might also want to decide on the appropriateness of adopting radical positions on the outcomes. An author may mistakenly think a conclusion becomes more acceptable because he or she ignores the centre-ground standpoint on the data and, instead, focuses only on the extreme perspectives.

Workbook

I i Abstract 1 (Worked sample appraisal)

1. **Read this abstract. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were addressing the researcher face-to-face (see below).**

Attend closely both to what is read *and* how this is communicated, and record thoughts immediately. Imagine we are being informed about the study in a conversation with the researcher. Feel free to interrupt and "say" whatever is necessary. Record agreement, disagreement, doubts, surprise, or even disbelief. Although our initial reactions may be countered, or opinions necessarily revised as a result of what is read later in the text, the idea here is to trigger a response in a number of ways, and so immediately begin to involve ourselves in the study. Look out for places where we think the meaning of something is unclear or incomplete, and then formulate a quick question to the researcher establishing what we would like to know to clarify things. Examine how the researcher uses empirical or intuitive knowledge to illustrate or defend something, and jot down a response to the point made. Be prepared to identify possible good or weak arguments or inconsistencies, concentrating particularly on the language used to present such points of view. At early stages in the paper, there will no doubt also be aspects of the procedures used about which we will want to see further details in later sections. Jot these down too, so that they can be referred to later on.

In this worked example, the columns have been filled in to show you one way of going about this initial reading task.

Using computer software to understand EFL writing and help teacher correction in large classes. Has a positive effect on production and gives information about errors.	The purpose of this paper is to study the use of a computer-based instrument to monitor, evaluate, and understand better ① EFL student writing in Country X. Specially-designed software, used together with software commercially available ②, was used initially to alleviate teacher obligation in the correction of written work ③ from the large numbers of students in Country X's EFL writing classes. It is suggested that objective analysis and feedback influence positively and significantly students' production, and also serve to present detailed information about the errors often found in particular writing genres.	*What do you mean by "use"? How can one program do all three things?* *But can a computer program do the same work as a teacher correcting?* *Do you mean "significantly" as a result of some statistical test on the data?*
Could help teachers in these classes. Test group errors significantly less than control.	I suggest that this kind of finding might help change the way teachers understand large class sizes in this country. Groups were divided into test and control ④ and they completed writing assignments during the first six months of 1995 using the software. *T*-test procedures revealed more statistically significant reductions ⑤ in the test groups' errors than in the control groups. I also describe other data which revealed the precise error types found throughout these writing genres.	*I don't see how this finding can bring about such an effect.* *So both groups used the software?*

2. **Read again the relevant section in the textbook introduction and then study my responses to the following questions:**

a. Can you see a clear statement of the topic and aim of the paper?
The topic of this paper appears to be computer-assisted language-learning. The aim of the study is less clear: I wonder how software such as this can be used to "monitor" or "evaluate" or "understand better" student writing, all of which would normally seem to require a more subjective judgement.

b. Is there a concise description of the sample and materials used?
No basic information is given as regards the subjects used. It appears they might have come from large classes in Country X, but some more information about who participated and the material used would have been useful here for me to judge the study's immediate relevance.

c. What details are provided about the procedures used and the way data were later analysed?
I am told that both groups used the software; how the groups then differed in the procedures used will hopefully be explained later. Data have been analysed using a procedure known as a t-test. I have made a note here to check in the relevant section to see what the specific test was.

d. Is there a brief summary of results, or the general trend of these, and are you told what conclusions are drawn from these?
There seems to be adequate information here about the results. However, the connection between the t-test results on different amounts of errors and the aims mentioned in the first sentence is not obvious. Similarly, I am not sure how the "other data" described in the last sentence will fit in with the original "purpose" stated. The conclusion given states "It is suggested that objective analysis and feedback influence positively and significantly students' production". I also made a note to check in the body of the text what "feedback" is actually being provided to the student by the software and how its positive influence was measured.

Observations

① What do you think of this as an objective of the study, and where would you look for more information to see if it was achieved?
I was not sure how the results of such a study might help us to "understand better EFL student writing". Results might describe the present situation and subjects, but I thought it sounded somewhat ambitious to hope for more generalisation or more insight than this — maybe the "Discussion" or "Conclusion" sections of the paper will throw more light on this.

② What information will you want to have about this software, and where would you expect to find it in the paper?
I got the impression here that ***two*** *kinds of software "treatment" (i.e., "commercially-available" and "specially-designed") were given to the two groups involved. I will be looking for more details of the software in the "Method and Procedures" section of the paper and also to see whether steps were taken, or needed, to measure or separate out the effects of each of the two kinds of software.*

③ What do you think is meant by "alleviate teacher obligation", and how could the software help?
If the alleviation of "teacher obligation" in correction involves what I currently

understand is the correction AND evaluation of written work, I cannot see how the objectivity normally found in computer software can hope to replace both these subjective teacher actions. I made a note to look out for, perhaps in the "Discussion" section, how the authors believe "teacher obligation" is still being alleviated if the final evaluation still remains in the hands of the teacher.

④ What else would you be looking to read about this division, and where would you most likely find this information?
I would be interested to see (in the "Method and Procedures" section) how the division was decided upon. Given the comparative aims here, it will also be useful to read there about whether these groups were considered equal to begin with.

⑤ What do you understand as yet by the term "statistically significant reductions"?
Presumably, this refers to significant differences in error frequency between the groups. It might have been useful to have included information here about the observed probability level ("p"). This could give me, as the interested reader, important information to judge whether the cut-off point reached in terms of significance is sufficient to warrant a more detailed reading of the "Results" section.

I i Abstract 2 (Guided appraisal)

1. Read this abstract. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were addressing the researcher face-to-face. Advice on how to go about reacting spontaneously to the text was provided on p. 153.

	A case-study ① was set up to investigate how far overt teaching of revision ② affects a group's written production and also the way they perceive the writing process. First, subjects, taken from two classes of High School Streams ③ in Country B, were taught to revise ④. This teaching took place once they had written a first draft of the composition. Also all the participants answered a specific set of questions before and after the study. A number of students were interviewed ⑤. An holistic measurement of performance ⑥ in writing tasks was made once at the beginning and once at the end of the research period ⑦ and the results compared with those subjects who were not taught revision strategies ⑧. A description is given of the nature of this instruction and results are reported on the effects of the overt teaching. It appears that this teaching did have a significant influence on production. After analysing the data it was seen that subjects varied as regards the way they thought about writing and revision⑨. I suggest that writing teachers might think about using the system of different drafts of a piece of writing instead of completing a piece of work in class, since results from this study indicate that overt teaching of revision can help students become more conscious of how foreign-language writing can be influenced by certain elements of the discourse itself.	

2. **Read the textbook introduction to this section again and then respond to these questions, using some of my prompts if you wish:**
   a. Can you see a clear statement of the topic and aim of the paper?
   b. Is there a concise description of the sample and materials used?
      Are there any possible implications of "subjects, taken from two classes…"? What information would you be looking for about subject selection?
   c. What details are provided about the procedures used and the way data were later analysed?
      Is there a noticeable difference between subjects being "taught to revise"

and also "taught revision strategies"? Where would you look in the paper for more information?

d. Is there a brief summary of results, or the general trend of these, and are you told what conclusions are drawn from these?
What do you understand here by "a significant influence on production" as a result of the teaching?

Observations

① Think about what could be meant by "case-study"?

② What do you understand by "overt teaching", and where would you expect to see this explained?

③ What more would you want to know about these "High School Streams", and where would you look for this information?

④ What information would you need about this teaching and the students' experience?

⑤ What would you need to know about this interview and those selected for it?

⑥ What differences might there be between "holistic" and any other kind of evaluation, and how might outcomes be affected here?

⑦ Think about the potential importance of the time-scale in this research and how this might affect the reliability of the measurement. Where would you look for more information, and what would you want to know?

⑧ What would you want to know about the two groups, and where would you expect to find this information?

⑨ What do you think this variation might likely consist of, and where would you find these details?

I ii The background to the problem and the problem statement 1 (Worked sample appraisal)

1. **Read ABSTRACT 1 again (p. 154).**

2. **Read this section below. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face.**

In this worked example, the columns have been filled in to show you one way of going about this initial reading task.

Origin of problem: large class sizes and exam demands makes it difficult to monitor students. Teachers need help.	What sparked off this research idea is the way foreign language is currently taught in Country X, where there are often very large groups of people in class, and teachers have a lot of work to do to get the students to the level required by external examinations. As a result, it has become difficult to establish the true level of skills each student has arrived at, where their weaknesses lie, and if they are improving to any extent. To make matters worse, teaching takes place in an educational context that places great emphasis on testing and objective skill measurement. Thus, it was felt that teaching staff faced by these kinds of classroom problems in Country X could benefit from some kind of instrument which provides an answer to this situation ①.	*Is it difficult because of the class size or the demands of the exam, or both?* *How will the instrument solve all these problems?*

Computer-based writing and correction program designed to help teachers save time in their correction but not replace them.	The answer was sought in a technological instrument that could help teachers in such a situation ②. This is a writing and correction system based on the use of a computer that makes use of currently-available technology and software programs, an example of which is *Grammatik* ©, and which are used in conjunction with our own software program ③. Our intention was not to invent a foolproof parsing machine or a better expert writing system, but rather an inexpensive instrument which would present no problems of introduction into the classroom for teachers in Country X. Similarly, we were concerned to design something that would not actually replace the teacher nor add to his or her current burdens. Rather, our aim has been to harness what we know about the positive uses of CALL ④ to a particular aspect of a normal writing class that often requires too much of the teacher's precious time ⑤.	*...so a number of programs are used as part of the software package?* *What problems might there have been in this introduction?*

3. **Read again the relevant section in the textbook introduction, and then study my responses to the following questions:**

1. Is the **background to the problem** described? If so, what is it?
   The author sets up the need for computer-assisted language learning through a "technological instrument" — to aid L2 writing/correction by referring to the current limitations and burdens placed on teachers as a result of the teaching context described.
2. Is there a **problem statement**? If so, what is it in your own words?
   This is not clear in this opening section. Implicitly, I have to assume that the study will describe the positive results obtained from the use of this technological instrument in this educational context. The abstract did provide somewhat more information about "feedback", but aims should also be made clear at this point.
3. From the **problem statement**, do you understand: (a) the variables to be measured? and (b) the functions of these variables? If not, what values would you assign from what you have been told so far?
   (a) and (b): Again, the lack of information within this section means I must return to the abstract to see what is being measured or contrasted here. Apparently, I am reading about a study that will compare two groups (of test and control) and that will demonstrate the advantages of this program. Perhaps the

dependent variable here will be something like "frequency of errors". Since both groups, according to the abstract, will use the special software, I am not yet sure what the function of the independent variable will be.

4. Is there a **contribution claimed to theory and to practice**?
It would appear from these introductory lines that the problem addressed here only has established contribution to practice. However, I assume that data here might also provide us with more information to add to the body of knowledge about computer-assisted language learning and, specifically, in the area of the teaching of L2 writing in such contexts.

The contribution to practice is evident: not only does the author see this instrument as providing "help [for teachers] in such a situation", but also an effort has been made to design a product that is inexpensive and easy-to-use, both of which may well be practical merits in such a teaching situation. What I am still unclear about is whether this instrument helps to relieve the teacher in some way directly (as seems to be implied in this section), or only indirectly, by helping the student solve their own problems of correction.

Observations

① Think about how, or whether, this program might realistically "provide an answer" to the situation described.
I already noted the somewhat ambitious claims in the abstract that the "computer-based instrument" would help to "monitor, evaluate, and understand better EFL writing in Country X". I still wonder how the use of only one technological instrument can help in a situation brought about by decisions made outside the classroom and probably out of the average class teacher's hands (i.e., the massification of classes and the concomitant requirement for individualised information on student progress).

② Why do you think the author considered "a technological instrument" would be helpful for these teachers?
Perhaps the teachers and/or students were already amenable to such technology, or perhaps the authors felt that such technology had intrinsic benefits to alleviate the problems in their current teaching situation.

③ What would you want to know about the way these programs were used together? Where will you look for this information? How will this be useful in your appraisal of results?
Since the experimental software is used "in conjunction" with other software,

I will look in the upcoming "Method and Procedures" section for more detailed information on just how the two programs differed, and, in particular, what specific needs the specially-designed software was designed to address that were not already tackled in the published program.

Since the software apparently consisted of a blend of programs, I would eventually be interested in discovering which elements of which program were thought to have had the best effect on results. As it stands, any significant improvement seen may have been the result of any combination of elements from all the software involved, rather than due solely to the authors' own specially-developed program.

④ Where would you expect to read more about these uses?
This statement will be best expanded upon and/or justified in the review of the literature to enable me to understand how the authors felt their own study would contribute to the established theoretical and practical advantages already assigned to computers in language learning.

⑤ To what extent do you think this program might have the potential to save the teacher time?
Little information is provided here about what elements of the writing class are actually taking up so much of the teacher's time and which are intended to be addressed by this program. I still wonder how much is gained eventually in terms of relieving teacher burden if — subsequently — the teacher still has to evaluate the student's writing, perhaps reviewing what has been corrected by the program and/or the way the student has interacted with the information provided by the program.

I ii The background to the problem and the problem statement 2 (Guided appraisal)

1. **Read ABSTRACT 2 again (p. 157).**

2. **Read this section below. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face. Advice on how to go about reacting spontaneously to the text was provided on p. 153.**

	For many researchers and teachers, student writers need to be able to learn more effective revision procedures ①; on the other hand, there is no firm agreement on the question of whether any significant progress can be made by direct teaching of task-based revision techniques. My basic assumption here in this study was that direct teaching of revision is feasible. An investigation was set up to determine the extent to which direct teaching of revision strategies affects both subjects' writing abilities and the way they see their writing ②. The basic objective set out was to study what effects such instruction had in the teaching context of Country B's secondary schools ③. I decided to experiment in this study with the use of direct teaching of revision subsequent to initial drafts and prior to the writing of final versions of compositions. Currently, little or no use is made of such multiple drafts in secondary schools in Country B. It may at first surprise that the practice of using single drafts lingers on in Country B, especially considering the swing in writing methodology over the past years from concentration on the one-version product to focus on process. At the moment, the normal requirement of an L2 student here is to write only a final version of up to ten or twelve compositions per year. Once compositions are corrected by faculty, these are returned so that the student may correct any grammatical errors. Any positive effects of the treatment here might signal the need to consider the introduction into schools' L2 writing curricula of a system of multiple drafts.	

3. **Read the textbook introduction to this section again and then respond to these questions, using some of my prompts if you wish:**

1. Is the **background to the problem** described? If so, what is it?
2. Is there a **problem statement**? If so, what is it in your own words?

3. From the **problem statement**, do you understand: (a) the variables to be measured? and (b) the functions of these variables? If not, what values would you assign from what you have been told so far?
   How many dependent variables appear to be involved? On what are the groups being compared here?
   Can you think of any control variables that might have been used in this design? If so, to what end?
4. Is there a **contribution claimed to theory and to practice**?

Observations

① How do you react to the expression "For many researchers and teachers…"?

② Think about the possible definitions of "writing abilities".

③ How realistic do you find this objective? What implications does it have for the research design?

I iii The review of the literature 1 (Worked sample appraisal)

1. Before you read this review of the literature, familiarise yourself with the study by reading again its accompanying "ABSTRACT 1" and "THE BACKGROUND TO THE PROBLEM AND THE PROBLEM STATEMENT 1" on pages 154 and 159.

2. Read this section below, written in 1996. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face.

In this worked example, the columns have been filled in to show you how you might go about doing this initial reading task.

Studies with native speakers have produced positive effects with computer-based feedback, but future work needs to concentrate on better design and other populations.	In general, there have been positive results ① from studies which have investigated the effect of computer-generated feedback on native English speakers (Researcher 1, 1983; Researcher 2, 1985; Researcher 3, 1987). Researcher 4 (1991) complains about the poor design used in these past studies ② and suggests that further research is needed before any firm conclusions can be made about the effects of this feedback. In general, we need better structure in our studies, with particular concentration on aspects of internal/external reliability and validity ③; such studies will also need to check on the effect of such feedback on other populations, particularly on ESL/EFL students over longer periods of study ④.	*"In general"? What have other studies discovered that was not positive?* *What has been concluded so far?* *Why do you insist on "longer periods of study"?*
There are positive effects of this technology in the classroom. Many programs exist to help writing.	Research has often reported that students are positively disposed towards using computer technology in the classroom and that such use in lessons does have a positive impact ⑤ (Researcher 5, 1989; Researcher 6, 1990; Researcher 7, 1986; Researcher 8, 1988; Researcher 9, 1990). There are a large number of computer programs commercially available which make claims such as this: "the easiest way to improve your writing" (found in the instructions to the program *Grammatik* ©). There has also been research carried out by specialists such as Researcher 10 (1989) who have tried to compare the way students write when they work with their tutors to when they work with programs that correct their grammar. It is surely not surprising that those using the computer program did not produce writing with the same quality as that produced by the group working with tutors.	*What specific "use" do you mean?* *Why is it not surprising? What would you have expected?*

Whether the teacher or computer helps more is a complex question; how and when are computers efficient? Local research reports generally positive results.	Many researchers have emphasised that there is more to the problem than just whether computers are better than teachers. Researcher 11 (1990) wonders how effective computer programs can be and whether their efficiency is the same at all stages of the writing process. Such questions need to be answered by the kind of detailed research recently done by Researcher 12 (1993a; 1993b) who found that using these programs in Country X produced overall improvements in students' attitudes and writing. Researchers 13 (the authors) (1994) report quite positively on the results of their customizing the program *Grammatik* © (which the authors also customized in the present study ⑥) to make it respond better to the particular needs of their students.	*Why compare the two anyway? Should it not be how computers can supplement, rather than replace the teacher?* *"quite positively"?: it sounds as if there were problems, too — what were they?*

3. **Read again the relevant section in the textbook introduction, and then study my responses to the following questions:**

Are you satisfied that the review (a) describes the most relevant work done and indicates its relative importance, (b) has sufficient critical address of the literature, (c) communicates the main points related both to the background to the problem and the problem statement/independent and dependent variables, (d) covers an adequate time-span, (e) has adequate reference, where necessary, to empirical work? In general, does it convince you of the need for the study?

a. *I had earlier assumed that the objective here was to "describe the positive results obtained from the use of this technological instrument in this educational context". Given this supposed objective, I wonder about the immediate relevance of including studies with native English speakers. Similarly, student attitudes to the software — to which a number of references are made in the second paragraph — does not seem to be part of the study. Finally, it would appear that the study wherein changes were made in the Grammatik© program is of far more direct relevance than others mentioned here, and more information about the methodology and/or data analysis may have been useful.*

   I find no obvious indication by the author that certain studies and their results have more bearing than others on the current investigation.

b. *The author has included a relatively large number of references in this review and from a variety of other researchers. However, the review is limited to the*

description of what were the summarised findings in each case. For example, the second paragraph opens with reference to five studies in order to support the summary statement that the interaction of student and learning context with computer technology has been positive. If this is considered of importance, it would have been useful to have had more critical address of these studies: did success depend on how the teaching was distributed within a lesson, for example, or do these authors recommend ways of improving student predisposition to the programs, which are now to be tested in this present study?

c. *Although the review does take in some studies on writing in the second paragraph, the first ones cited appear to have concentrated on native English "speakers", and it is not clear whether the objective was to study their written output or their overall language learning. There is a discernible movement towards the problem statement in that I then begin to read about the effect of specific programs on writing. The second paragraph cites what seems to be a similar case to this in which test and control groups were set up to test the effect of the software. However, little detail is given to enable us to appreciate how the present study makes an advance on what was revealed. The final paragraph further narrows the focus onto the present context in the same country (Researcher 12) and another study which adapted the same computer software (Researcher 13); however, where we would now expect to read more important comparative information from this same teaching context or research area, tantalisingly little is forthcoming.*

My recap of the paragraphs highlighted a lack of continuity or logical flow in the review. For example, I did not find it easy to follow the internal logic or see the overall aim of the second paragraph, moving as it does from describing reactions to software to how certain software is marketed. Similarly, there would seem to be no obvious connection between the first and second paragraphs. The general impression given me is of a number of somewhat randomly-connected aspects of using computer-generated software eventually rather loosely focussing on the problem statement.

d. *Given that the paper was written in 1996, I would like to have been informed why work carried out so many years earlier is still considered relevant here (e.g., Researchers 1 and 2 (1983) and (1985) or Researcher 7 (1986)). If this work was to be deemed "classic" in the field, then this could be indicated or highlighted by the author, so that I am aware of the central relevance of such studies, despite their age.*

e. *Most of the work here seems to have been empirically based. It would, perhaps, have been useful to have the specific contexts of the empirical work in*

references to Researchers 10 and 13. This would also enable me to weigh up the relative importance of such previous study to the current work carried out in Country X.

On the whole, the literature review does not convince me of the need for this study in the current context. Given the potential importance of the contribution and application of "specially-designed software" to the problem of large classes, I was surprised to find only one fleeting reference (Researcher 12) to relevant work on this application. More critical engagement with the literature cited might have helped to highlight the current gaps in our knowledge that make it necessary or advisable to carry out the present study in this particular context. The fact that other researchers think more research needs to be done may not, in itself, be enough to justify the study.

Observations

① What might be understood by the term "in general"?
This gives me the impression that I am going to be given some idea of both positive and negative findings. Although the target group was "native English speakers", perhaps some information could have been given about some of the conflicting findings hinted at here, if only with which to compare results obtained in this study.

② What might be the implication of this remark for the present study?
My reaction here was to assume that the poor construction of previous studies is, in some way, to be improved on in the present study. By embarking upon such improvements, the author is building up part of the contribution to the field from his or her own study. Thus, some critical engagement here with some of the specific weaknesses mentioned would have been useful.

③ What might this refer to, and where would you expect to read more information?
Reliability is not normally referred to as "internal or external" although validity is discussed in this way. I made a mental note to look out for more specific information in the upcoming "Method and Procedures" section. I assume that such previously-detected problems have been addressed in the present study. However, if other studies did present problems of internal and/or external validity, this could seriously affect the amount of confidence to be placed in their results.

④ What is your reaction to this remark in the context of the present study?
The implication is that this study will indeed be "over longer periods". It appears from the abstract that the study will be over a six-month period. However, I still do not know to what extent such a time period will, in fact, be an improvement on these previous studies. I also do not see that a case has been made for the need for work with EFL/ESL students based on the literature reviewed so far.

⑤ What could be understood here by "positive impact"?
A large number of studies is cited, and some indication of the nature of this impact could usefully have been given.

⑥ What more would the reader want to know about the authors' actions?
Since the use of this program also appears to have been central to the objective of the present study, the reader might benefit from having more information about the kind of modifications (and their outcomes) carried out in other key studies.

I iii The review of the literature 2 (Guided appraisal)

1. Before you read this review of the literature, familiarise yourself with the study by reading again its accompanying "ABSTRACT 2" and "THE BACKGROUND TO THE PROBLEM AND THE PROBLEM STATEMENT 2" on pages 157 and 163.

2. Read this section below, written in 1998. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face. Advice on how to go about reacting spontaneously to the text was provided on p. 153.

A common idea throughout the recent literature is that revision strategies may not have received sufficient attention from teachers in the classroom. There is an observation by Researcher 1 (1995), for example, who feels that teachers tend to spend more time reading about such strategies than actually teaching them. Moreover, there has also been debate as to whether the explicit teaching of revision is pedagogically useful or advisable. One of the reasons for this is that revising may easily be understood by students to be a "quick-fix" solution when writing compositions. Results from many studies have shown students revising at surface text level ①, and more often than not failing to address revision of meaning at all. In the face of so much evidence that these writers prefer to avoid revision beyond the surface level, it has been suggested that such concerns might be better addressed, and with more effectiveness, before pen is set to paper (Researcher 2, 1993). Textual revision may also take in other areas, such as discourse-related awareness of reader, along with the aim of the writing and its internal coherence and flow.

The teaching of revision has been approached in many American college classrooms as a peer work activity with a collaborative approach, precisely to avoid seeing revision as part of a prescriptive process. Here, such peer work and group conferences are the usual way revision is taught. It is also possible to use activities which involve the whole class, and a number of tasks might be employed which actively exploit the use of the teacher at the front of a whole class, such as a combined critique of a particular text (Researchers 3 and 4, 1996). Researchers 5 and 6 (1990) and Researcher 7 (1986) have also described other activities which help students of a foreign language revise more efficiently ②.

On the other hand, we do not read of conclusive findings with regard to the success of instruction in revision in L1. Researcher 8 (1982), for example, showed that explicit teaching can result in improved performance ③ of school students, but Researcher 9 (1978) reported no significant changes in the writing performance of thirteen college writers after direct teaching of revision strategies. Conversely, other researchers have found that less than ten minutes of direct teaching can produce significant progress in the second draft of compositions compared to a control group who had not received any instruction at all (Researchers 10 and 11, 1991).

Those who have compared L1 and L2 revision have discovered that the two processes are analogous (Researcher 12, 1990; Researcher 13, 1986); it follows that we have sufficient information to enable us to study how far direct teaching of revision strategies affects L2 writing proficiency ④. The little research that has been conducted up to now suggests that L2 revision does not always produce improvement in writing. Researcher 13 (1995), for example, discovered that revision actually led to more errors being made in writing and also had a negative effect on writing anxiety. Researcher 14 (1996) reported on a study with unsuccessful EFL student writers and showed that they were unable to revise for meaning. He further suggested that these learners needed formal instruction in revision. The teaching profession itself has also joined the call for more studies into revision strategies (Researcher 15, 1995; Researchers 15 and 16, 1994). The problematical connection between exercising revision in an L2 text and seeing improvement was also reported by this author in 1998 ⑤.

	Finally, studies have revealed that encouraging students to write over a number of drafts is not often used in schools in Country B and that the normal practice is for schools to apply prescriptive rules for the way L2 writing is to be produced (Researcher 17, 1991; Researchers 18 and 19, 1995). Typical practice is that writing staff are told to ask for a minimum number of class compositions and, as an integral part of staff inspection, these compositions are regularly checked by the Head of Department ⑥.	

3. **Read the textbook introduction to this section again and then respond to these questions, using some of my prompts if you wish:**

Are you satisfied that the review (a) describes the most relevant work done and indicates its relative importance, (b) has sufficient critical address of the literature, (c) communicates the main points related both to the background to the problem and the problem statement/independent and dependent variables, (d) covers an adequate time-span, (e) has adequate reference, where necessary, to empirical work? In general, does it convince you of the need for the study?

a. *How relevant to this paper are studies about revising without specific instruction, and those referring to one specific country?*
   How relatively important are studies carried out in the same country as this present study?
b. *Do you read of any apparently conflicting or controversial results here that seem to require more critical address?*
c.
d.
e.

Observations

① What more might you want to learn about these studies?

② What else might you be interested to learn here?

③ Consider what could be meant by "performance" in this context.

④ Comment on the logic in this sentence.

⑤ What is your reaction to this comment on the author's previous finding?

⑥ Consider the possible local consequences of such procedures for the application of any successful outcomes from the present study.

I iv Research questions and hypotheses, variables, and operational definitions 1 (Worked sample appraisal)

1. Before you read this section of the paper, you might like to familiarise yourself with the study by reading again its accompanying "ABSTRACT 1", "THE BACKGROUND TO THE PROBLEM AND THE PROBLEM STATEMENT 1" and the "THE REVIEW OF THE LITERATURE 1" on pages 154, 159, and 165 respectively.

2. Read this section below. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face.

In this worked example, the columns have been filled in to show you how you might go about doing this initial reading task.

Students' skills will improve as a result of using this computer program and error profiles will become available.	The research hypothesis in this study was that objective measurements and feedback will have a significantly positive effect on students' skills. Furthermore, these measurements and feedback will also enable detailed profiles to be made of the errors common to specific writing genres.	*Why didn't you opt for a null hypothesis? What do you mean by "students' skills"?* *"...will also enable..." — is this also part of the hypothesis?*

3. Read again the relevant section in the textbook introduction, and then study my responses to the following questions:

1. (a) Are research questions or research hypotheses formulated? If so, what are they? (b) Are the research questions exploratory, descriptive, or explanatory? (c) Are the hypotheses offered directional, and do they predict differences or relationships between variables? (d) Are the research questions/

hypotheses unambiguous, consistent with the problem statement, feasible, and supported by the review of the literature?

a. *A research hypothesis is described which, as it stands, indicates that two variables (i.e., "objective measurements" and "feedback") will affect "students' skills" and also give detailed information about specific student errors. It is not clear to me whether the latter is meant to be part of the hypothesis or, in fact, an additional research question.*

b. *Not applicable here.*

c. *The tone of "…will have a significantly positive effect on…" indicates a directional hypothesis that looks towards some measurable difference between the groups and improvement in these skills. As was mentioned in the textbook, directional hypotheses should ideally be seen to proceed logically from what we have been told about previous studies in the literature review section of the paper. In this case, I noted that the first part of the review did indeed highlight positive outcomes of computer-generated feedback, despite some perceived research design faults. However, I was told such positive results and attitudes had been obtained with "native English speakers" rather than the present population. On the other hand, we are later told of previous studies in Country X that "produced overall improvements in students' attitudes and writing". Although I will need to understand what the "skills" mentioned actually are, this would seem — in theory at least — an hypothesis that can be tested.*

   As regards alternative relationships possible here, I suggest the relationship between using a software program offering feedback and seeing improvement could be more complex (and less easily observed and measured) than it initially appears in the hypothesis. It might be argued that success might also come as a consequence of the way one chooses to work with the software, rather than the software as such. As the literature review indicated, improvement may also be associated with an individual's attitude towards the technology. Perhaps improvement could also come about as a result of a more subtle — and less observable — combination of software content, familiarity with using computers to write assignments, and the ease with which the effect of the software can be integrated into the individual writing process.

d. *I was faced with having to read between the lines in what is a somewhat imprecise hypothesis. In my own words, the researcher thinks that "independent evaluative data and feedback will help to improve students' skills and that these will also help to give us exact descriptions of the kind of*

typical errors made when these students write in a particular style".

Firstly, I am not sure which of the two variables (or the two jointly) is being seen as predicted to bring about the desired effect. Secondly, I could not see how something like "objective measurements" can effect changes in skills. Thirdly, such a variable may well be manipulated and expected to affect student skills, but it would not be considered a normal direct property of such a variable also to "enable detailed profiles to be made of the errors ".

As regards consistency, the problem statement left me in doubt about the real purpose of the study — whether it would be a descriptive study of the "technological instrument" or a comparative study looking at the effects of the instrument using control groups — although the abstract had spoken of "test and control…groups" (page 154). Also, I remember that the problem statement talked of developing a "writing and correction system". There might be some inconsistency here, as I felt that providing "feedback" on skills requires something more than merely "correcting" that production.

With regard to the feasibility of time engagement here, the researcher had already warned in the literature review section of the paper about the need for more attention to "longer periods of study" in research studies investigating computer-generated feedback. I felt that subjects will need to be in regular (and perhaps prolonged) contact with the software for any "significantly positive effects" on student production to be reliably registered. I will be interested to see how improvement is measured, and how much is considered relevant: a long-enough period with any new learning device will usually conclude with some positive effects on learning being seen.

Since the review of the literature in this paper was seen to provide a certain amount of positive support for the use of computer-generated feedback (albeit mainly from L1 studies), it seems logical that an hypothesis that basically posits the same outcome is supported to a certain extent. On the other hand, there would appear to have been no support — beyond intuition — for the second "part" of the hypothesis that states such computer-generated activity will provide detailed information about student errors.

2. Can you identify the **principal variables** of the study, and are these to be measured as **nominal, ordinal, or interval scales**? Comment on the perceived appropriateness of these scales. Are moderator or control variables evident?

This would appear to be a study wherein effect is anticipated on a dependent variable.

Reading directly from the hypothesis, the dependent variable here appears to be "students' skills". The paper so far seems to indicate that what is being investigated are students' skills in correcting their written errors rather than all or any other skills. The abstract did mention use of "a test group….and a control group" and "reductions in the test group's errors", so I assumed that some measurement will be made of these "errors" at some time to permit such a comparison. The independent variable will be used to effect this comparison, but I am still not sure what is being manipulated here. There is no evidence here of any explicit moderator or control variables. No indication has so far been given whether subjects are to be sub-categorised into, say, levels of proficiency for the purposes of the study.

There is as yet no indication of the proposed measurement of the variables. However, I read in the abstract that "t-test procedures" were carried out, and I assumed that continuous (most probably, interval) data are going to be forthcoming from the variable "students' skills".

3. Can you predict any **intervening variables** or contributory factors — if not stated here — that might affect findings?
   I hypothesised a number of factors that might intervene in any direct cause-effect link between the independent and dependent variables in this study. Attitudes (towards using the computer and writing), ability (to use the computer), or familiarity (with the machine) may not be directly measurable, yet may well play some role in affecting expected research outcomes. Similarly, sex and language proficiency level of subjects might also affect outcomes, although no work is cited in the literature review.

4. What were the **constructs** used and have these been adequately delineated to permit operational definition? How have these constructs then been **defined operationally**, where necessary, and is this description acceptable in its present form?
   There are no operational definitions provided as yet for "students' skills" or for "objective measurements and feedback". I also do not understand what "significantly positive effect" will mean. Also, as it stands, "errors" could be understood at best as any category, or quality, of errors. This may not be exclusive enough to permit the kind of specificity mentioned as desirable in the textbook section. I made a note to look for more details in the following "Method and Procedures" section.

I iv Research questions and hypotheses, variables, and operational definitions 2 (Guided appraisal)

1. Before you read this section of the paper, you might like to familiarise yourself with the study by reading again its accompanying "ABSTRACT 2", "THE BACKGROUND TO THE PROBLEM AND THE PROBLEM STATEMENT 2" and the "THE REVIEW OF THE LITERATURE 2" on pages 157, 163, and 170 respectively.

2. Read this section below. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face. Advice on how to go about reacting spontaneously to the text was provided on p. 153.

	Two research questions were investigated: 1. Does overt teaching of revision strategies bring about greater improvement in written production than the traditional way of teaching revision? 2. Does this teaching of revision strategies have any effect on the way students perceive the writing and revision process, and if so, how? [*From "Method and Procedures" section* ……I decided on the method and content of the revision strategy teaching together with the teacher involved. We agreed detailed revision teaching plans which would aim to stimulate audience needs/reader awareness in the writers and would focus on three main areas: Evaluating, detecting, and repairing problems. We further agreed on a focus of teaching which encouraged students to read each others' work and through which the students might appreciate how their writing could be made easier to read by concentrating on a text's appropriateness of style and good organisation of information.	

	The control group, who received the traditional teaching, had input for their compositions before starting to write ①. They were told to complete their compositions in class ② and were given a little help in the form of teacher correction of their surface mistakes. Improvement in writing was measured as the difference between the pre- and post-test marks assigned to the student. To discover students' views on writing and revision, students in the two revision groups were asked to complete questionnaires in class before and after the study. Subjects in the control group completed a similar version with fewer questions ③. I also carried out a number of semi-structured interviews with some of the subjects from one of the experimental groups ④ which helped provide further insights into the influence of this teaching on students' views of the writing and revision process.]	

3. **Read the textbook introduction to this section again and then respond to these questions, using some of my prompts if you wish:**

1. (a) Are research questions or research hypotheses formulated? If so, what are they? (b) Are the research questions exploratory, descriptive, or explanatory? (c) Are the hypotheses offered directional, and do they predict differences or relationships between variables? (d) Are the research questions/hypotheses unambiguous, consistent with the problem statement, feasible, and supported by the review of the literature?
   a.
   b. *Do you feel the researcher is aiming to build or to test an hypothesis here?*
   c.
   d. *Has the previous literature review adequately supported these questions as regards the local teaching situation?*
2. Can you identify the **principal variables** of the study, and are these to be measured as **nominal, ordinal, or interval scales**? Comment on the perceived appropriateness of these scales. Are moderator or control variables evident?
3. Can you predict any **intervening variables** or contributory factors — if not stated here — that might affect findings?

What other local factors might affect students' perceptions about the revision and writing process?

4. What were the **constructs** used and have these been adequately delineated to permit operational definition? How have these constructs then been **defined operationally**, where necessary, and is this description acceptable in its present form?

Observations

① What more would you want to know about this input, and where would you look for this information?

② Can you think of any possible conditions, and consequences, of writing "in class" which might affect outcomes here?

③ What is your reaction to this information in the light of the second research question?

④ Consider the possible reasons for, and consequences of, this selection of subjects.

II i Subjects and materials 1 (Worked sample appraisal)

1. **Before you read this section of the paper, you might like to familiarise yourself with the study by reading again its accompanying "ABSTRACT 1", "THE BACKGROUND TO THE PROBLEM AND THE PROBLEM STATEMENT 1", "THE REVIEW OF THE LITERATURE 1", and "RESEARCH QUESTIONS AND HYPOTHESES, VARIABLES, AND OPERATIONAL DEFINITIONS 1" on pages 154, 159, 165, and 173 respectively.**

2. **Read this section below. As you are reading, in the *right-hand column*, write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face.**[1]

In this worked example, the column has been filled in to show you how you might go about doing this initial reading task.

---

1. Since this, and the following "Results", section typically present summarised details of procedures or findings, it is not considered necessary for readers further to summarise what they read here.

Over three months ① during the autumn term of 1994, eight classes were chosen to take part in the study. The total number of subjects involved was 374 students, all of whom came from two higher education establishments ② in Country X. Of these eight classes, six were final-year students at the Country's only Institute of Business Studies and were in the EEC Business and International Accounting Departments. The remaining two classes consisted of second-year students from the main university of the capital, who were reading Business Studies.	*"chosen": who did the choosing, and based on what criteria?* *So they were all studying Business Studies? What would be the implications for any eventual generalisation of results? Were they comparable in other senses?*
The subjects had to write a series of detailed business letters ③ with the help ④ of my specially-designed software. This software formed part of a larger package and had been especially adapted to the needs of my students ⑤. The program functions as a text editor and provides advice for the students as they are working on how to write business letters. The program also has the advantage of being able to store the errors made during its use and to form databases of the types of errors made during each letter-writing task. Subjects were told at the start of each session the kind of letter to be written and the context of the same. Five different types of letter were required: an application, an inquiry, a response, a sales letter, and an offer.	*Were they used to doing this? And with this software?* *"My"? Which students? These present or previous ones? It might make a difference.* *I'd want to know more about how this advice works on screen…*

3. **Read again the relevant section in the textbook introduction and then study my responses to the following questions:**

1. What basic identification data are provided about the **subjects**, and are these data sufficient to permit replication?
   Subjects are identified as coming from the local university and business institute; I learn of their respective courses and their current year of study. In terms of replication, however, much more needs to be known and much of this could affect my understanding of the eventual findings.

   Subject age and L2 proficiency level might well be seen as important factors in a study that is implicitly looking at the way subjects respond to computer-assisted language learning. This information is supplied only indirectly here: I can only guess at the age of "final-year" and "second-year" students based on my own experience, which may or may not reflect the situation in Country X. Adequate replication of this study would require some basic information about

L2 abilities. If I assume there were some basic differences by virtue of academic year and/or institution, I would need to treat any generalised findings with considerable caution. After all, to what L2 proficiency levels can we apply any outcomes?

Finally, in order to provide for adequate replication, a study of student/computer interaction should surely have provided information about the kind of computer experience subjects have both on a personal and class basis. Indeed, in my comments on the research hypothesis, I suggested that subjects might stand a greater chance of obtaining better results with new computer programs if they were already familiar with their operation and, secondly, if they were already used to managing computers in their L1 and/or L2 writing.

2. a. What are your initial reactions to **the numbers involved or any grouping** envisaged?
   b. Do these **groups** reflect the original pre-group sample in terms of their basic characteristics and is any justification provided for the eventual group size?
   a. *I read that 374 subjects are to form the sample, although I do not know whether a larger sample was initially used to arrive at this final total after further selection. Since eight classes are involved, this would represent an average of over 45 subjects per class. Each reader would have their own understanding of whether this is a large number of students in class. Furthermore, whether this is a representative large class size of the population to which the findings are directed can only be assumed here, as I am not given any further information. Representativeness would also include the composition of the groups in terms of gender and nationality. I assume that all group members were nationals of Country X. As regards gender, there are no details and — again — I do not know whether the final sample reflects what is normally found in large classes in Country X. More seriously, perhaps, only the total size of the final group is given; it is impossible to work out whether, for example, individual classes in one institution consisted of greater numbers than in the other institution. It seems the business studies institute contributed considerably more subjects to the study than did the university. There is no reason given for this.*
   b. *I have yet to read about how the control and experimental groups described in the abstract are to be set up. However, I would need to be looking for clear criteria for grouping, not least because the initial sample seems to be so large and heterogeneous. Similarly, I do not have enough information to know how far the control and experimental groups will*

reflect the original pre-group sample. The few details of basic identification provided and the imbalance built into the initial selection of intact classes suggest this may be difficult to obtain.

3. Can you see any potential threats to internal validity of the data from **attrition**, **history**, or **maturation** factors?
   I know from the abstract that the study took place over "six months". There is a large number of subjects from different classes and courses, and different colleges. It is not, as yet, clear how the researcher actually managed such numbers throughout the six-month period, but it is reasonable to suppose that there is potential for attrition in this sample; this could, for example, potentially affect any posterior statistical comparison if there was considerable attrition in one of the eight classes involved.

   I found myself wondering about how to manage so many students in different contexts throughout the period of treatment and ensure that the only, or main, feedback on written errors was indeed coming from the computer software. In particular, I wondered how far these subjects' own institutions were also concurrently providing/demanding EFL writing practice as part of their specific courses. Indeed, given the nature of these, such writing might well have consisted of the very kind of business letter-writing required in the present study.

4. What information is presented concerning the way subjects were **selected** and/or group membership assigned? What do you see as the consequences of this as regards eventual **generalisation of findings**?
   Previous appraisal did note a tendency in this paper to wish to generalise any findings here to the wider context of Country X. The first thing I notice here is that intact classes have been used; indeed, the researcher uses the word "chosen", which may indicate some other specific and un-stated criterion behind the selection. I am not told so far of any attempt to randomise the group assignments in order to reduce this initial selection bias. Secondly, subjects are all chosen from one main study area: Business Studies. Logically, therefore, any findings would only be generalisable to such groups. I have yet to see any conclusions in this regard, but this would already seem to clash with the wider objective put forward in the abstract: "to help to monitor, evaluate, and understand better EFL writing in Country X".

5. Has any **material or instrument of testing/measurement** been satisfactorily described and/or samples provided? Where appropriate, has its development/design and scoring been adequately discussed?

Despite the central role played by the software in this study, little information is as yet provided about it and the way it is used. From the description given, it seems only to function very much like a normal word-processing program ("a text editor") which incorporates an error-correction complement. What inherent qualities show it has been "specially-designed" for the particular "needs of my students" is not clear, although I note that it has a capacity to record errors made. No details are provided about how these errors are to be scored.

I made a note in the analysis of the abstract of the paper to check here for more information about how this software might "alleviate teacher obligation in the correction of written work". However, how it is able to provide "detailed information about…errors", give help to teachers, and give "advice for the students" on their writing is not clarified. I also noted in the abstract that this software was to be used "together with software commercially available" — later described as the program "Grammatik©" in the literature review. Although this is indeed a commercially-available program, some basic identifying information would have been useful here for those not familiar with the program itself and interested in using it in their own studies.

6. For any instrument of testing or measurement used (including observation), what evidence of **reliability** was given, and how acceptable is this evidence?
   Although not, strictly speaking, a test instrument, there is still no specific evidence produced for the "reliability" of the software used. Some informal support in this regard could have been provided (although it provides no direct proof of reliability) by describing in more detail the development and previous use of similar programs in Country X mentioned in the literature review.

7. For any instrument of testing or measurement used, what evidence of **validity** was given, and how acceptable is this evidence? If none is given, what do you consider to be possible threats to validity here?
   Too little information is provided here accurately to judge questions of validity. I did read in the abstract, however, that the aim of the study was to "study the use of a computer-based instrument to monitor, evaluate, and understand better EFL student writing (my underlining)". *I find myself wondering how valid data will be in this case, since the software is being used here with students studying English for a specific purpose (i.e., business), and who are being asked to write what we are told are "…detailed business letters".*

8. How have the **main independent and dependent variables** been realised within the method itself, and how satisfactory do you find this?
   Variable assignment has been uncertain throughout this paper so far. I previously described the dependent variable here as "students' skills", which had been elsewhere understood to refer to the outcomes registered through their writing. Data for this variable now appear to be coming from the five different letters the subjects will produce and the number and type of errors registered by the program. The independent variable has been realised as the use of the specially-designed software package; I am still not sure how groups will differ on this. There is still no explicit identification of any control or moderating variables, and no information has been forthcoming concerning the possible influence of the intervening variables or other factors I suggested in the previous analysis (such as computer experience or L2 proficiency).

Observations

① Why do you think three months were needed?
It is not clear why selection should take place over such a relatively long period. If this corresponds to some specific selection procedure or any difficulties encountered, I would have been interested to learn about it for replication purposes. For example, it could be that this time was needed to contact and solicit the agreement of institutions and/or subjects or to choose between a number of different classes, based on some kind of criterion.

② What more information would you be looking for about these institutions, and why?
Both for replication purposes and for questions of data validity, I would be interested to know why only two institutions were used, if these were also "chosen", and whether these institutions are representative of similar large-class institutions in Country X. Similarly, it would be useful to have some information about the kinds of courses and L2/EFL teaching they undertake.

③ Why would it be important to learn about the subjects' experience writing in these genres?
If students are not familiar with this kind of writing in their own institutions, it might be argued that the task itself presented certain difficulties for these subjects, and that therefore results are also affected by this, rather than just as a direct outcome of the treatment (or lack of it).

④ What might you want to know about this "help"?
Some sentences below I also read of "advice" provided by the software: I would be interested to read whether such help or advice was with reference only to the errors made by the subject or — as it seems from this paragraph — consisted of more wide-ranging information about writing, such as content, style, and so on. If this were the case, it would be useful to know how this actually worked in practice with the subject writing at the computer. Similarly, are errors automatically corrected by the computer program, or is merely advice given without any suggestions for correction possibilities? If the former, how far would the end result then be a valid representation of the "students' skills" as such, and how far a result based on the "combined efforts" of the student and the advice from the computer?

⑤ What do you understand by "my students" in this context?
I am not sure whether the researcher is referring to the fact that the software was designed for "his" or "her" own students (i.e., previous to the study) or to these present subjects. If the former is the case, one wonders how far the present subjects (and the class sizes involved) represent the same population — with the same needs — for whom the software was originally "specially-designed". If they do not, there might be a question about the intrinsic value of data from subjects for whom the software was not actually intended.

II i Subjects and materials 2 (Guided appraisal)

1. Before you read this section of the paper, you might like to familiarise yourself with the study by reading again its accompanying "ABSTRACT 2", "THE BACKGROUND TO THE PROBLEM AND THE PROBLEM STATEMENT 2", "THE REVIEW OF THE LITERATURE 2", and "RESEARCH QUESTIONS AND HYPOTHESES, VARIABLES, AND OPERATIONAL DEFINITIONS 2" on pages 157, 163, 170, and 177 respectively.

2. Read this section below. As you are reading, in the *right-hand column*, write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face. Advice on how to go about reacting spontaneously to the text was provided on p. 153.

One high school was selected for the study and four fourth-year (of five) classes chosen. Since local education authorities would not agree to random selection of either school or classes, these four classes were decided on by the principal of the centre, a school which in turn had been allocated by the head of the local education authority. All the class members were female and between fifteen and sixteen years old. The family backgrounds of these children were varied, with an overall rich mixture of middle- and working-class families ①.

From these four original classes, the principal and two of the school's English teachers proceeded to select two which would receive the revision strategy treatment. The subjects in these two treatment groups were then informed by their teachers what different kind of teaching they would be given ②. The third class was only to be used to demonstrate the improvement in writing performance. The fourth class was not involved for comparison purposes since students there were mainly studying History of the English Language as a subject and so merely received some extra hours of English per week.

Teaching staff were the following: during the first six months of the study one of the groups who received revision strategy instruction was taught by a local (non-native) teacher, who also gave classes to the third (control) class. A native English teacher taught the other revision group. However, late into the study, this latter teacher left the school and — in the last three months — his teaching was taken on by two new teachers. Such changes were not thought to have been a serious threat to the findings since these teachers continued with the same revision teaching procedures.

Classes in the school are typically large, with over forty pupils in each. Although pupils talk to each other in their native language, the majority of core subjects in the school are taught in English. The normal teaching procedures used in the school with regard to EFL writing were the following: very little direct teaching of such writing goes on at this level and class time is usually taken up with timed compositions in class and/or doing the corrections to the same. School policy requires teachers to give back compositions with any errors clearly marked, and students then have to correct the grammatical mistakes underlined. Students are normally expected to present for the Cambridge First Certificate examination near the end of their studies, where the composition plays a major role; therefore, much of the EFL writing centres on learning how to present acceptable writing of this kind with few grammatical mistakes, but content is not considered such an important feature.

As regards the materials used, all subjects were given the following written instructions for their writing task before the study began and then again a year afterwards, after the end of the research period: "Write a composition on the following, giving reasons for your arguments: "Smoking should be banned in all public places" Do you agree?" ③.

These pre- and post-test compositions were given to two examiners who were not taking part in the study. They handled all the compositions from both experimental and control groups between them. These examiners used the same guidelines and composition marking scale as that used by Cambridge First Certificate examiners. No pre-marking coordination meeting was held between them as they were both experienced examiners ④. The writers' names were whitened out of the papers before these were marked, and examiners were not told which compositions came from which group nor which sitting (i.e., before or after the study). An inter-rater reliability coefficient was calculated after the marking was completed and considered satisfactory at .89 for the compositions written before the study began and .78 for post-study writing.

Interview data were collected as the study was nearing its end ⑤. Eleven students from one of the revision classes had an interview, the aim of which was to obtain more personal accounts of what these subjects thought about revision and the teaching they had received. A pool of subjects was made (n = 18), based on those that appeared to have made the best and the least improvement during the year, and each subject approached. Eleven volunteered to be interviewed by the researcher. Semi-structured interviews were recorded and aimed to collect data about what subjects saw revision to be about, if they thought it was important, and the importance they attached to the reader ⑥. All interviews were tape-recorded and subjects' replies analysed and collated.

Student opinion about the teaching received was also collected from questionnaires (see sample, below) which were given out before and after the research to those subjects in the two groups receiving the revision strategy instruction (see below). As subjects were obliged to complete these questionnaires in class, there was a 100 per cent return rate on these. The aim was to see the effect of the revision teaching received on the way these subjects saw the writing process. Questions were divided into sections, which concentrated on: what they thought would constitute a good piece of writing, what they liked and disliked in the special teaching they had received, and what they thought revision is all about. The control group had to fill in a similar version of the questionnaire but which had fewer questions under each sub-section ⑦.

EXTRACT FROM QUESTIONNAIRE FOR EXPERIMENTAL CLASS
NAME: .....

I want to know what you think about the writing and revision classes you have had.

1. Now that you know how to revise, are you happier writing your compositions? Yes/No
2. Now that you know how to revise, do you think you write more effectively? Yes/No

See how far you agree or disagree about the following statements:

3. i. Learning how to revise has been a great help for me. Yes/No
   ii. Revising is not very interesting for me. Yes/No
   iii. Learning how to think about the reader of my writing was a great help for me. Yes/No
   iv. Learning different ways to make a plan of my composition was a great help for me. Yes/No
   v. Learning how to evaluate someone else's composition was very helpful. Yes/No
   vi. These classes will help me when I take the Cambridge First Certificate. Yes/No
   vii. I liked the idea of my friend checking my own composition. Yes/No
   viii. I will be able to use the strategies I have learnt on my own from now on. Yes/No
   .....
4. **Circle ONE of the following options as the most important according to your point of view.**
   i. "For a good composition, you need to have..."
   interesting content
   good paragraph structure
   few grammar mistakes
   ii. "For a composition to be organised correctly, it should have..."
   a topic sentence
   five paragraphs
   spaces between paragraphs
   an introduction and conclusion
   iii. "Revising a composition correctly means.....".
   correcting any grammar mistakes
   reading through the composition and changing any confusing content
   reading through the composition, thinking of the reader, and changing things as a result
   checking for, and changing, any spelling mistakes

3. **Read the textbook introduction to this section again and then respond to these questions, using some of my prompts if you wish:**

1. What basic identification data are provided about the **subjects**, and are these data sufficient to permit replication?
   How similar do you think the groups might have been? Is it important?
   What specific information would you want to have about L2 teaching of writing in the school? Is enough given here?

2. a. What are your initial reactions to any **groupings** envisaged?
   b. Do these **groups** reflect the original pre-group sample in terms of their basic characteristics, and is any justification provided for the eventual group size?
   a. *What information would it be useful to have regarding these four intact classes?*
   b. *How might the discarding of the fourth group affect the final data obtained?*
3. Can you see any potential threats to internal validity of the data from **attrition**, **history**, or **maturation** factors?
   How might the parallel teaching of other subjects affect results here? What observations might be made concerning the teachers involved in the study?
4. What information is presented concerning the way subjects were **selected** and/or group membership assigned? What do you see as the consequences of this as regards eventual **generalisation of findings**?
   What observations could you make about the selection for the interview?
5. Has any **material or instrument of testing/measurement** been satisfactorily described and/or samples provided? Where appropriate, has its development/design and scoring been adequately discussed?
   Comment on any other material (i.e., apart from the questionnaire and interview schedule) that appears to have been used.
6. For any instrument of testing or measurement used (including observation), what evidence of **reliability** was given, and how acceptable is this evidence?
   Comment on the perceived precision of the questions in the questionnaire and the fact that they are presented in the L2.
   What do you think might be gained or lost by eliciting responses using closed (yes/no) questions and multiple choice items?
   How do you respond to the fact that questionnaire sheets asked for subjects' names?
   What informal elements of the marking may increase our confidence in the reliability here?
7. For any instrument of testing or measurement used, what evidence of **validity** was given, and how acceptable is this evidence? If none is given, what do you consider to be possible threats to validity here?
   Judging from this extract, how far do you think this questionnaire might succeed in gathering accurate/valid data about "the effect of the revision teaching……on the way these subjects saw the writing process"? And what about the data from the interview?

What more might you want to know about the "guidelines and composition marking scale" used, given the first research question (see p.177)?

8. How have the **main independent and dependent variables** been realised within the method itself, and how satisfactory do you find this?

Observations

① How might socio-economic background have affected results?

② Consider the possible consequences of this information given the subjects.

③ What other information about the conditions for writing might be relevant here?

④ Examiners of…?

⑤ Compare, and comment on, the period of data collection from the interview and the questionnaire.

⑥ How might the demands of these subjects' current learning context also have affected these opinions?

⑦ Given the research questions, why do you think such data were gathered from the control group?

II ii/iii Procedures, research design, and data analysis 1 (Worked sample appraisal)

1. Before working through this section of the paper, you should re-read the corresponding "SUBJECTS AND MATERIALS" text and appraisal on pp. 180–185.

2. Read this section below. As you are reading, in the *right-hand column*, write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face.

In this worked example, the column has been filled in to show you how you might go about doing this initial reading task.

The letters, once written, were given to the teachers as hard copy and on diskette. The hard copy versions were evaluated in the usual fashion, but the computer versions were passed through the special software program to reveal specific errors. The program is designed to monitor a total of 45 error types drawn from a database of the most common errors of EFL intermediate-level ①② students in Country X when writing ③. The list of errors noted was then given to the student responsible for the writing.	*When were they handed in? Immediately? Might this not have affected their corrections? And what do you mean by "usual fashion"?* *How did the program "monitor" the errors?* *...but then what was he/she supposed to do with it?*
Grading for both groups took into account the errors found by the program in the sense that the more errors that were revealed, the lower the final grade awarded. However, only the subjects in the experimental groups were permitted to see the specific errors in their returned writing since they received the computer error printouts together with the personal feedback from the teacher. The control groups only received the latter feedback.	*"took into account"? So it was the only criterion used?* *How was this scored? — did it only depend on the frequency of errors?*
Comprehensive descriptions of subjects' errors were obtained during the six months of the study ④ using the specially-designed software and combining its use ⑤ with Grammatik as the parsing tool. In order to evaluate the extent to which feedback generated by the computer impacted on the specific errors, two control groups were chosen from those studying final year at the Institute of Business Studies. One group was formed from the subjects in the Business department and the other from the International Accounting department.	*What did these "descriptions" consist of? Do you mean "records"?*
All groups were using the same computer equipment; this was decided in order to control for any effects that might have been due to the positive feelings these subjects had towards using the computer rather than due to the software itself. The experiment was designed in such a way as to ensure that the majority of factors would be consistent across all the groups. Again, it would thereby be clearer that any differences between the groups would have been caused by use of the software program.	*What equipment did they use, and did you check for their computer skills beforehand?* *What factors do you mean?* *Isn't "caused" rather strong?*

3. **Read again the relevant section in the textbook introduction and then study my responses to the following questions:**

1. What is your appraisal of the **timing of events**, and is this information sufficient to permit **replication**?
   I previously noted the longitudinal nature of this study and questioned how improvement was to be measured across the research period. In this kind of study, where the cumulative effects on L2 error correction of interaction with specific software appear to be a focus of attention, it is going to be of interest to the reader to understand how the nature of this interaction might have affected results throughout the six months of study. However, I am not informed about the period of time elapsing between each of the five letter-writing sessions. Equally crucial is information about how these individual writing sessions were divided up: how long were subjects given to plan their work?; did they then have to begin writing immediately?; was this against the clock?; were subjects given specific time to check their work, and could the test group see the computer feedback on their previous letters when they were writing?; and when did these subjects get the printed feedback?
2. Are there any potential threats to internal validity as a result of **test or practice effect**?
   Although no formal pre- or post-tests as such are carried out with the subjects, I wonder how far the cumulative practice obtained as a result of the five letter-writing sessions and subsequent feedback affects the results obtained. In other words, the experimental group can be hypothesised constantly to have had the "advantage", not only of the software and the feedback itself, but also of the continuous use of this.
3. What is your assessment of any **instructions** given the subjects?
   In the first of the two sections of the "Method and Procedures" reviewed, there was some implied summary of formal instructions, although I was not told the form (i.e., written or oral) or the language in which these were presented. I presume these would have needed to have been quite specific on the day, however, since subjects in both groups would have had to know not only what they had to write, but also how much time they had to write, how this might be divided up, how many words were required, and so on. It might be argued that the ability to correct written mistakes on the computer may be related as much to such demands of the immediate writing context as to the software itself. Since these compositions were later evaluated and returned to subjects, I also presume that these were told to identify themselves on their texts. I wondered whether subjects, instructed to identify themselves and conscious of the fact that their writing is subject to posterior evaluation, might not be more

attentive to the surface correctness of that writing than they might be in other anonymous situations.

4. What potential threats of **reactivity** do you see with respect to i. observer/scorer effects and subject expectancies, and ii. observer/scorer bias?
   i. *Since I am not told what information about the study was in the hands of the subjects, it is impossible to judge whether experimental group members might have been reacting to the investigation rather than the treatment itself. I note, however, that six of the eight groups chosen were studying in intact classes at the Business Studies Institute and that two control groups were chosen from amongst them. It is possible, therefore, that the experimental groups in the same Institute were aware of their "special status" (or became so, after receiving the special feedback) and that this could have given them an extra impetus to do well. Even if we assume that information would not have leaked about the study's objectives or the differences in feedback afforded both groups, it might well have become clear to either group from the evaluations that "the more errors that were revealed, the lower the final grade awarded" (p. 192); as a result, both groups might have become additionally motivated to check for errors carefully. I wonder if this potential contamination could have been avoided by allocating control status to the two remaining classes studying at the main university.*
   ii. *Although scoring is done mechanically, human bias may be introduced through the "personal…teacher feedback" afforded both groups. I am not told how many teachers were involved in the marking, whether these were the teachers normally responsible for the classes, nor whether these were aware of the study and/or the groups to which each student belonged. What is now clear is that this feedback was additional to that provided by the computer in the case of the experimental group. Although not directly reflected in the score awarded, it is conceivable that this too played its part in any subsequent improvement. The lack of information about the evaluation instrument and the instructions given the teachers also means that I do not know how far this feedback might have been affected by the tiredness and boredom potentially induced by marking so many compositions. The nature of teacher feedback would need to have been coordinated beforehand to ensure that certain subjects did not have an unfair "advantage" in their feedback as a result of any particular detail obtained therein.*
5. i. What details are provided about the **environmental conditions** of the study and could these have affected outcomes?

ii. What observations about **external validity** can be made in the light of these details?

i. *I am not told here the conditions in which the writing was undertaken, nor whether this was the same for all the subjects involved. However, I note that all groups used the same computer equipment. It is not clear from this how far all subjects were actually used to writing in this way and in these conditions in their normal classes. Writing at the computer might not yet be a typical L2 writing experience for many outside this environment and this would need to be taken into account in the interpretation of any results here. That is, how far is what we see in this study a direct consequence of the specific writing conditions, rather than the software feedback (or lack of it) itself?*

ii. *The considerations in i. above will obviously mean that the reader would need to consider the extent to which such conditions for L2 writing in this experiment are likely to be similar throughout Country X. Similarly, the lack of information about how time was divided up in the sessions makes it impossible to judge the external validity of any time limits.*

6. In the light of what you have read in this section, do you wish to amend, or add to, your previous comments on **group assignment**, **materials**, and the potential threats to internal validity of the data from **attrition**, **history**, or **maturation**?

*That generalisation will not be advisable in this study has now been further emphasised by the assignment of control/experimental group status: I am told that both control groups were "chosen", rather than randomly assigned. No reasons are given for this choice and the apparent imbalance it creates between the two sets of groups (cf., question 4i). Also, no further evidence has been provided to demonstrate that the groups were equal before any treatment was given and comparison made. Indeed, we read that "the experiment was designed in such a way as to ensure that the majority of factors would be consistent across all the groups (*my underlining*)". In the light of such an observation, I might need to question the validity of any comparisons between groups after treatment.*

I remain unaware of the instructions given the subjects, and it is also not clear how the experimental material (i.e., the computer feedback) was to be used here. There seems to be potential for the experimental subjects to be using the feedback however they wish. This may eventually present me (and the researcher) with a problem of interpretation: how do we know if success is down to the software itself or to the way the individual has interacted with this?

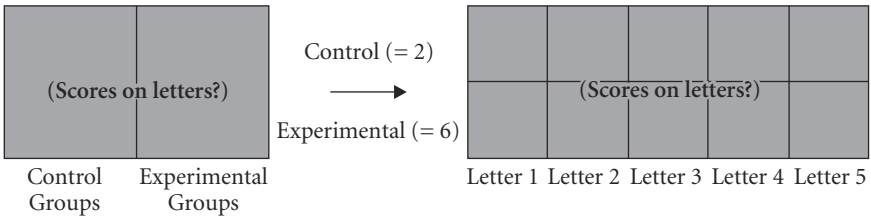
I do read that this program monitors for up to 45 "common errors" and, presumably, identifies these in the feedback. However, no further information is provided to respond to my previous question about scoring: how are the letters to be scored, given the fact that t-tests were to be applied to what should be continuous score data?

No information on attrition is provided, and I assume that all 374 subjects continued on through the six-month period of the study. There is no further light thrown on the question of whether the researcher considered parallel college class activities as potentially impinging on outcomes here.

7. Identify the **basic type of design** employed here and **draw the design box**. What immediate observations can you make about this design and its consequences for the study?
There are eight groups, two of which have been assigned control status and the rest, apparently, experimental status. There has been no random selection of subjects (intact classes are used) nor random assignment to these groups. Furthermore, there has been no attempt to pre-test the groups for initial equality before applying the treatment. This is the kind of design described on p. 75, which straddles a pre- and a quasi-experimental procedure.

This seems to be a mixed design with respect to subject comparisons, since independent groups are involved (eight different classes), but these same subjects are to assessed and followed across five different letter-writing assignments, making this also (potentially, at least) a repeated-measures design. Any information yet to be confirmed has been designated with a question mark to remind myself to re-consider this, when appraising the appropriateness of the design for any posterior data analysis:

Main research hypothesis: "Objective measurements and feedback will have a significantly positive effect on students' skills"
Independent variable: Writing Group
Level 1: + software feedback (Experimental)
Level 2: − software feedback (Control)
Dependent variable: Scores on letters(?)



My attention is drawn to the resulting imbalance built in to the division of control and experimental status (i.e., six in one and two in the other). I have made a note to see how the researcher presents and analyses such data in the upcoming "Results" section. There is also a manifest possibility for data-gathering and comparison here at different stages (i.e., at each letter writing).

8. Attempt visually to **classify the data-collection procedure,** and comment on the perceived consequences of this for any eventual findings. Where necessary, suggest how this might have been improved, and why.
No pre-testing was undertaken with the two groups. Intact classes are used ("chosen"), but I am not told whether or not they were equal to start off with. However, at this point, a number of the questions already raised about the present design of the study hinder my ability to classify the procedure more accurately. For example, while a "treatment" as such is involved, it has not been made clear how this is expected to be used, or how it is expected to reflect on posterior measurements of writing. The basic idea seems to be the following:

$$\left.\begin{array}{l} O(1) \xrightarrow{?} X(1) \longrightarrow ?\,O(2) \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ O(1) \xrightarrow{?} X(0) \longrightarrow ?\,O(2) \end{array}\right\} \text{etc. up to } O(5)$$

I have also noted the problem of interpretation presented by the fact that the control group are <u>also</u> receiving (teacher) feedback of an unspecified nature (X(0)). Since the idea may be to compare the groups at each letter-writing session (O), the design indicates that appropriate concurrent measurement will not be straightforward. For example, at letter number two (O(2)): subjects in both groups would presumably have received their particular feedback (X 0/1) from letter number one at some time before this measurement, and this is hypothesised as showing (or not) an effect now in this letter. However, I do not know how long the teacher or computer feedback has been in the hands of the respective groups, what they are told to do with this, nor the time periods between each measurement. In this scenario, if both groups are measured at the

same time, it might indeed tell us whether either group is ostensibly getting better or worse in their written outcomes, but will probably tell us little about the true comparative effect of the feedback as such. Testing for differences only at the end of the five-letter writing period might be equally inconclusive.

Improvements could be suggested, firstly, by random allocation of groups, as well as some kind of pre-test. Some consideration could also be given to matching students across groups as a result of this test, to enable a more precise comparison of treatment effect to be made.

Another interesting alternative would have been some type of time-series design (see p. 75). A group from only one institution could be followed over the six-month period: the "treatment" (X) would be applied once a number of pre-tests were made and over a longer period and then the normal behaviour of the group established. This would solve one of the big problems in the present design: we do not know how the groups differ at the start. A further "security device" that might be implemented would be to use a similar control group across the same time period and compare the two:

O(1)-O(2)-O(3)-O(4)-X-O(5)-O(6)-O(7)-O(8)

O(1)-O(2)-O(3)-O(4)-O(5)-O(6)-O(7)-O(8)

9. What **procedures are identified for data analysis**, and do these **deal adequately with the original objectives of the study**? In the absence of information about procedures, suggest how this might be done.
The only information so far revealed (in the abstract) about data analysis has informed me that "t-test procedures" were used to reveal "more statistically significant reductions in the test groups' errors than in the control groups". Since t-tests are normally used to reveal differences between two levels or groups of one independent variable, this would seem initially to be reasonable here (but see below, questions 10 and 11). It remains to be seen whether the t-test used is parametric or non-parametric.

A look back to the research hypothesis shows that two objectives were presented: the effect of feedback on students' (writing) skills (presumably addressed by the t-test analysis) and "detailed profiles to be made of the errors common to specific writing genres". This last would seem to be addressed by the program itself which, together with the parsing tool, is able to identify and track 45 special errors and provide "comprehensive descriptions" of these. What these "profiles" actually are, and how the program goes about creating them, is not clear, but the implication so far is that little more than counting

will take place. I have made a note here to follow up on these profiles in the "Results" section.

10. Provide a **step-by-step description of the elements involved in the data analysis** so far, and decide on the appropriateness of any proposed analysis procedures in the light of this.

Variables: *Dependent* = Scores on letters; Measurement = scores? (interval?)
Independent = Groups (two levels); Measurement = nominal (experimental and control).
Comparison or relationship to be tested: Comparison between independent groups (between-groups). The possibility also exists for comparison within the same groups (repeated-measures) on the five assignments.

From the flow chart, I follow through the information provided: that the researcher wishes to "discover the effect of an independent variable on a dependent variable", that this variable is measured (apparently) as a test ("scored") measurement, that the main objective sees different groups being compared, and that two levels are involved in the independent variable. The t-test procedure is thereby confirmed as applicable here, and I have made a note to check in the next section on whether the researcher has opted for a parametric or non-parametric test. On the face of it, given the limitations of the design, it would seem that the researcher would be wiser to opt for a potentially less powerful (i.e., non-parametric) procedure.

I do not yet know whether the researcher is interested in further comparison within (or across) the groups on the five assignments, but such comparison would suggest the option of using a 2 (group) by 5 (letter) ANOVA matrix.

11. Have the **necessary assumptions** associated with the stated or implied analysis procedure been met in a way that suggests the reader can have confidence in the results of the analysis?
The basic t-test has four assumptions assigned: independence of groups, independence of observations, normality, and equal variances. As regards independence of groups, no information was provided to suggest that the experimental and control classes in the Institute of Business Studies — from which six of the eight classes were chosen — could not have exchanged information about the study. Large groups of subjects were deliberately used here, of course, to satisfy the objectives of the particular study; however, I wondered whether cross-attendance at classes was adequately controlled.

There does, indeed, seem to have been independence of observations, since scores would appear to be based on the computer tally of errors (see p. 192) and not on the personal teacher assessment/feedback. Certainly the numbers involved are enough to suggest a normal distribution might be obtained. I might expect both the mean and s.d. to be appropriate measures for the upcoming t-test. However, the researcher seems to have implied that results from this study will be used to apply to large classes of "EFL…writers in Country X": I wonder how representative of such classes are 374 subjects studying Business Studies in "two higher education establishments". Finally, a formal appraisal of equal-variance will need to wait for the s.d. measure in the "Results" section. I note at this point, however, that the n size of the experimental and control groups are most certainly not equal here (i.e., two control groups vs. six experimental groups) and that, therefore, this assumption might be violated.

Observations

① What might you want to know about the scoring procedures for these types?
I will be interested to see an itemisation of these types so that some idea can be formed about the kind of error, or absence of error, that was scored in each observation. I read elsewhere here that the more errors made, the lower was the score awarded. I wondered what scale was used to decide on the final score, what happened when an error was repeated throughout a letter, and whether this was "scored" as one or more errors in this grade. Equally, did the program consider all 45 errors of equal gravity for the final score? For example, it could be argued that errors of punctuation or capitalisation are of less import in this business-letter-writing context than, say, spelling or tone of address and that this should be reflected in the scoring procedures.

② How appropriate is this particular database here, considering the proficiency levels involved?
The problem here is one of comparing like with like. My question is how far this instrument can be considered valid with these subjects; they do all come from Country X as far as I know, but I have been given no information about their EFL proficiency levels and, therefore, whether these error types can be expected adequately to reflect the kind of mistakes the present subjects typically make.

③ How appropriate might this particular database be when used with this kind of writing?

The above observation leads me to wonder how far an EFL-based instrument like this might appropriately be applied to what, after all, appears to be ESP-based writing. Again, I would need to have been told more about the kind of writing carried out to build the database, but "general" EFL students would almost certainly be undertaking different kinds of writing to the present subjects with their business studies backgrounds.

④ How might the length of time of the study affect results of using the software?
I am not sure whether the six-month period was imposed by the context of the research, or whether this was deliberately chosen by the researcher as a period within which some kind of effect was to be noted. The literature review did suggest that research "over longer periods of study" was needed, but no reasons were given for this. The amount of time with which subjects are actively involved with the software (or its feedback) could prove crucial to its effectiveness.

⑤ Comment on the possible consequences of combining two programs in the treatment.
I still wonder whether the researcher can point with confidence to any positive effects on error correction as being a direct result of the specially-designed software (feedback) only. There might be a rather more subtle combination of factors at work, such as the ease with which the subject copes with the task itself, the interaction of the subject with the computer, his or her ability to spot and correct error on the screen and, now, the combined effect of both these software programs used.

II ii/iii Procedures, research design, and data analysis 2 (Guided appraisal)

1. Before working through this section of the paper, you should re-read the corresponding "SUBJECTS AND MATERIALS" text on pp. 186–189.

2. Read this section below. As you are reading, in the *right-hand column*, write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face. Advice on how to go about reacting spontaneously to the text was provided on p. 153.

Before and after the study, all the participating groups provided a sample of their writing in order to get data on writing proficiency. Improvement in writing was measured as the difference between these pre- and post-test marks assigned to the student. They also completed a questionnaire. For the six compositions written by the experimental group and the twelve written by the control group ① the procedures adopted in the experimental and control classes were the same.

Each class lasted 75 minutes. In the first lesson the experimental (revision strategy) group were given certain pre-composition input. They then started the composition in class and returned a first draft of the same some time later ②. The control group would follow the traditional mode whereby they also received the input but started and finished the composition in class. In the second class, the experimental group were handed back their first drafts with teacher feedback for perusal and were given initial revision strategy information and practice; the control group received their marked compositions back and were given a short time to correct these. They then also received the input, and then started and finished the next composition in class. In the third class, the experimental group received more detailed input and practice on revision and applied this to their own and a partner's composition. They then started another draft of the same composition, which was given in the following week. The control group were given their marked writing back and were again given time to correct the mistakes ③. In the fourth class, the experimental class received comments on their final versions and pre-composition input on the next composition. Again they started the next composition in class and returned a first draft of the same some time later. The control group also received the input but started and finished the composition in class.

Subjects and titles of the compositions had to be agreed with the class teachers responsible and the principal of the centre, and there was constant interaction between these throughout the study. We agreed detailed revision teaching plans④ which would aim to stimulate audience needs/reader awareness in the writers and would focus on three main areas: Evaluating, detecting, and repairing problems. We further agreed on a focus of teaching through which the students might appreciate how their writing could be made easier to read by concentrating on a text's appropriateness of style and good organisation of information.	

3. **Read the textbook introduction to this section again and then respond to these questions, using some of my prompts if you wish:**

1. What is your appraisal of the **timing of events**, and is this information sufficient to permit **replication**?
   What further information about timing of events for each group both within and across classes might have been useful for the reader?
   How might the timing of the interview/questionnaire data collection (see p. 186) possibly affect the data obtained?
2. Are there any potential threats to internal validity as a result of **test or practice effect**?
3. What is your assessment of any **instructions** given the subjects?
   Comment also on the instructions given subjects for filling in the questionnaire. (see p. 178 and p. 188)
4. What potential threats of **reactivity** do you see with respect to i. observer/scorer effects and subject expectancies, and ii. observer/scorer bias?
   ii. *Comment also on how potential bias was or was not controlled for the evaluation of the interview/questionnaire data. (see p. 178 and p. 188)*
5. i. What details are provided about the **environmental conditions** of the study and could these have affected outcomes?
   ii. What observations about **external validity** can be made in the light of these details?
6. In the light of what you have read in this section, do you wish to amend, or add to, your previous comments on **group assignment**, **materials**, and the potential threats to internal validity of the data from **attrition**, **history**, or **maturation**?

7. Identify the **basic type of design** employed here and **draw the design box.** What immediate observations can you make about this design and its consequences for the study?
8. Attempt visually to **classify the data-collection procedure,** and comment on the perceived consequences of this for any eventual findings. Where necessary, suggest how this might have been improved, and why.
   What purpose(s) does the pre-test appear to serve here?
   How might matched subjects provide an alternative approach to this design?
   How might this design also provide a further source of interesting data?:

   X(1)-O(1)-X(0)-O(2)-X(1)-O(3)-X(0)-O(4) → etc.

   where X(1) is the experimental treatment and X(0) is some other kind of "treatment"?
9. What **procedures are identified for data analysis**, and do these **deal adequately with the original objectives of the study**? In the absence of information about procedures, suggest how this might be done.
   The second research question suggests the possibility that the teaching of strategies has somehow affected opinions of writing and revision: how far do these procedures promise adequate comparative pre-/post data?
10. Provide a **step-by-step description of the elements involved in the data analysis** so far, and decide on the appropriateness of any proposed analysis procedures in the light of this.
11. Have the **necessary assumptions** associated with the stated or implied analysis procedure been met in a way that suggests the reader can have confidence in the results of the analysis?

Observations

① Comment on the possible consequences of the different amounts of writing completed by both groups.

② Comment on the possible consequences for outcomes of the different conditions for writing and revision experienced by the experimental group.

③ Comment on the differences in the distinct class procedures for the control group.

④ Do you think these obliged consultations and "constant interaction" might affect outcomes? If so, how?

III Results: The presentation and nature of findings 1 (Worked sample appraisal)

1. Before working through this section of the paper, you should re-read the corresponding "RESEARCH QUESTIONS AND HYPOTHESES, VARIABLES, AND OPERATIONAL DEFINITIONS", "SUBJECTS AND MATERIALS", and "PROCEDURES, RESEARCH DESIGN, AND DATA ANALYSIS" texts.

2. Read this section below. As you are reading, in the *right-hand column*, write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face.

In this worked example, the column has been filled in to show you how you might go about doing this initial reading task.

Table 1 shows the total percentage of errors made by all the groups across the five different letters. Only those errors that comprised 5% or more of the total are shown.	*So a lot of errors (of the other 38 categories) were under 5%, it seems?*
Table 1. Total errors across classes and assignments	
Category type — %	
Spelling — 27	*That seems a very large proportion of errors for "Spelling". I wonder why?*
Typical — 18	
Noun Phrase — 10	
Punctuation — 9	
Sentence variety — 7	*What does "Sentence variety" include as errors?*
Subject/Verb Agreement — 7	
Capitalisation — 5	
Those errors classified as "typical" formed a special category that was based on a data-base of over two hundred errors ① that this researcher had collected during his marking of compositions over the previous two years in Country X.	*But did these come from the same kind of writers and writing as here?*
The next table presents the most frequently-encountered error types (>5%) in the control group for each of the five letters and serves to show how error occurrence changed across this writing.	*Why should I only be interested in the most often-encountered errors in the control group only?*

Table 2. Most common error types across five letters for control group
A – Application; B – Inquiry; C – Response; D – Sales; E – Offer

A	B	C	D	E
Spelling	Spelling	Spelling	Spelling	Spelling
Punctuation	Typical	Typical	Typical	Noun phrase
Sentence variety	Noun phrase	Capitalisation	Noun phrase	Sentence variety
Typical	Subject/Verb agreement	Noun phrase	Subject/Verb agreement	Subject/Verb agreement
Subject/Verb agreement	Sentence variety	Punctuation	Punctuation	Punctuation
Noun phrase	Capitalisation	Capitalisation	Sentence variety	Capitalisation

Are these in order — top most common to bottom least common?

I noticed that the number of errors decreased at a significant rate for both groups the more letters written and the more time went on. In other words, the genre of letter required did not affect this decline. Having said this, the reduction in error rate computed for the experimental groups was greater (−1.14) than the control group. The following five tables show the mean error computations for the two groups across the five letters. Only those differences that turned out to be significant have been highlighted ($p < .10$).

What do you mean by "at a significant rate"? And did the order in which the letters were completed correspond to any pre-determined criterion?

Table 3. Letter 1: Mean errors of control and experimental groups

Error Type	Experimental Group Mean	Control Group Mean
Adverb ②	.35*	.12
Article	.38*	.41
Spelling	4.2*	2.3
Typical	1.7*	1.1
Redundancy	.32*	.05
Sentence variety	1.3*	1.7
Poor adverbs	.02*	.3

*$p < .10$

That seems like quite a low probability level. Is that the alpha level you chose?

Graphs would have helped me to understand all this better.

Table 4. Letter 2: Mean errors of control and experimental groups

Error Type	Experimental Group Mean	Control Group Mean
Adverb	.13*	.33
Infinitives/Gerunds	.11*	.19
Noun phrase	.67*	1.2
Punctuation	.68*	.34
Spelling	1.5 *	2.6
Typical	1.2 *	1.8
Verb object	.31*	.18
Too-long sentences	2.3 *	2.9
Sentence variety	.5 *	.31

*$p < .10$

Table 5. Letter 3: Mean errors of control and experimental groups

Error Type	Experimental Group Mean	Control Group Mean
Adjective	.37*	.23
Poor commas	.14*	.45
Incomplete sentences	.35*	.46
Verb object	.43*	.26
Prepositions	.79*	1.3
Poor pronouns	.96*	1.3
Relative pronouns	.79*	1.6
Subject/Verb agr.	.77*	1.3
Spelling	2.1*	2.6
Typical	2.5*	2.1
Abbreviations	.44*	.29
Question-formation	.68*	.42
Split infinitives	.21*	.43

*$p < .10$

Table 6. Letter 4: Mean errors of control and experimental groups

Error Type	Experimental Group Mean	Control Group Mean
Subordination	.41*	.67
Adverbs	.57*	.98
Verb forms	1.1*	1.7
Tenses	.52*	.86
Infinitives/Gerunds	.66*	.51
Subject/Verb agr.	.62*	.84
Spelling	2.2*	3.1
Punctuation	.24*	.43
Capitalisation	.22*	.49
Slang	.19*	.43

*$p < .10$

Table 7. Letter 5: Mean errors of control and experimental groups

Error Type	Experimental Group Mean	Control Group Mean
Article	.14*	.39
Sentence connections	.36*	.85
Poor pronouns	.43*	.96
Spelling	.89*	1.4
Punctuation	.52*	.32
Split infinitives	.31*	.11

*$p < .10$

It is hardly surprising that the first letter brought about such a relatively large amount of errors. Feedback had yet to be introduced to the group. More interesting outcomes can be studied by examining results of the subsequent letters (i.e., once feedback might be hypothesised to be having some effect). In order to see the amount of change that took place for the occurrence of each error type, the error mean committed by the experimental group was subtracted from that of the control group and the total result totalled for letters two, three, four, and five and for all the error types. Thus, for example, if the control group mean for "Punctuation" was totalled as 2.39 across the four letters and that of the experimental group 2.87, the sum computed would have been −.48. Such a negative number would have been an example of an error type that occurred at a higher rate in the experimental group than in the control group. Conversely, positive results would indicate the opposite. These computations for the 45 error types showed that the experimental group's errors were less than the control group's for 32 types. Of these positive changes, the greatest improvements were found in "Spelling" (2.4), "Subject/Verb agreement" (.6), "Noun phrase" (.4), "Verb object" (.3), "Verb form" (.2), and "Sentence connectors" (.1). Conversely, "Punctuation" (−.3), "Sentence variety" (−.2), "Possessives" (−.2), and "Typical" (−.1) types registered higher rates in the experimental group. These results therefore show the kind of error that is susceptible to improvement using the software program ③. By far, the "Spelling" type is seen to show the most sensitivity to improvement with the program. This was also where both groups committed most errors and was probably the easiest error to spot. The negative outcome for "Punctuation" is also interesting; the experimental group committed more errors than the control group. This might be because correct punctuation requires more profound knowledge of grammar than the computer can provide. The advantage for the teacher here is that spelling appears to respond well to such computer feedback and so the teacher would not need to spend so much time correcting poor spelling ④. On the other hand, large amounts of punctuation error give the teacher the chance to dedicate more time in class to this kind of problem.

How did you judge when the feedback might begin to show some significant effect?

How do you know what individual use each student made of the feedback given? Was this monitored?

So what is gained if the teacher ends up having to dedicate more teaching time to something?

3. **Read again the relevant section in the textbook introduction and then study my responses to the following questions:**

1. Does your initial reading of this section suggest that enough data have been provided so as to have **adequately responded to the research questions or hypotheses** previously put forward?
   The original research hypothesis/question suggested that "objective measurements" and "feedback" from the software would positively affect "students' skills" (later qualified as "errors"), and that detailed information would become available about specific errors as a result of the data obtained. The tone suggested a directional hypothesis wherein some improvement would be seen in the relevant group. Furthermore, I thought the word "significantly" might indicate that results will satisfy a certain cut-off level of hypothesised probability (i.e., the alpha level) in the t-test.

   A first reading suggests that comparative data have been provided for the two groups involved and that these could show whether there was improvement as a result of the use of the program. A rather "generous" significance level ($p<.10$) has been provided, although no (alpha) cut-off point was suggested prior to these results. I also wondered how the feedback provided might give the "detailed profiles of…errors common to specific writing genres" promised in the second part of the research question/hypothesis. These "profiles" are presumably displayed in Tables 3–7 (see below).

2. What **tables or graphical displays** of results are provided, and what do you understand from the data displayed? Are there any data that you feel might have usefully been added to the information provided here?
   I would have also been interested (Table 1) in seeing how many of the original 45 error types observed by the software were identified. It appears the remaining 38 types registered such low occurrence as to be of little interest to the researcher or the objectives of the study. If "detailed profiles" of errors are to be presented, it would probably be just as important to discuss both why certain error types occurred so often <u>and</u> why others were much less frequent.

   The total percentage covered by these seven types is 83%, which means that the remaining 38 types accounted for only 17% of the remaining deviance. Perhaps some types even failed to register any error occurrences. This seems to present me with a very small number of error types accounting for a considerable amount of error across a large number of subjects ($N=374$). If, indeed, this data-base derives from "…the <u>most common errors</u> of EFL intermediate-level students in Country X when writing" (my underlining), *I wonder if these*

present writers are so typical of this population.

In Table 2, I am not told how results are ranked and/or what the interval score is that divided the ranks. The researcher claims that the information provided "serves to show how error occurrence changed across the writing". After the initial total group information about frequency of error types, does the ranked information about error types from only the control group add anything more to my knowledge of the success of the software feedback program? Similarly, I am not sure what is meant by "control group", since there were two classes involved. Is this ranking based on the mean data of the two?

I suggested the writing these students might have been doing as part of the normal business studies courses could have impinged on outcomes here. Thus, it would have been useful to see how far the frequency of error occurrence changed across both control groups, even though the overall ranking might have remained the same or very similar throughout the research period. My worries about any threats to attrition across the long (six-month) study period could have been allayed by providing details of the subject numbers involved at each assignment.

As the "detailed…profiles" promised in the research hypothesis, the information in Tables 3–7 seems to be somewhat deficient. Only significant differences have been presented, but I would also have been interested in those error types that did not achieve significance or that narrowly missed the cut-off point — perhaps to see where the software feedback did not help so much. Groups were significantly different on only a relatively small number of the 45 types. Not all the significant differences revealed are "in favour of" the experimental group. I note — in the third letter — that the latter registered significantly more errors in "Adjective", "Verb object", "Typical", "Abbreviations", and "Question-formation" categories. Indeed, differences still exist after the final letter in "Punctuation" and "Split infinitives" categories. Might this, again, be evidence that the software program seems to help improvement in some, but not all, error types?

The probability level "$p<.10$" may be somewhat too low for me to have great confidence in the success of the software. However, as no alpha level was previously determined, I do not know if this probability was calculated as a post-hoc result after seeing the data or whether this represents, in fact, the (alpha) level of confidence the researcher had originally set for the results.

3. What information is provided by any descriptive statistics about the **distribution of data**?

The mean is used in the data as the measure of central tendency. However,

there is no measure of variability reported. Consequently, there is no way the reader can know what the spread of scores away from these means actually looked like. Working backwards from the demands of the announced t-test, I have to assume that the s.d. was actually calculated to be used subsequently in the formula to obtain the t value. However, by not reporting the s.d., and by only reporting some of the means, it becomes impossible for me to consider the assumption of normal distribution of data here.

4. Have the data **been scored using the unit measurement predicted** earlier, and/or has any appropriate **data conversion** taken place?
   Up to now in the paper there has been little clear information about the way error performance is to be scored. I was told only that "…the more errors that were revealed, the lower the final grade awarded". Presumably, some conversion of data must have gone on to convert the frequency of "the errors found by the program" (see p. 191 "Procedures, Research Design, and Data Analysis 1") to the kind of "grade" found in Tables 3–7. What the range of possible scores/grades is, or how this conversion was carried out, remains unknown.

5. i. What, if any, **specific statistical operations or calculations** were carried out on these data, and does this seem to have been carried out appropriately?
   ii. In the light of what you have read in this section, do you wish to amend or add to your previous appraisal of the **assumptions met** for this procedure?

   i. *The only specific naming of any statistical procedure to be used was in the abstract to this paper, where I read "t-test procedures revealed more statistically significant reductions in the test group's errors than in the control group". The particular version of the t-test used is not identified, and I cannot confirm its parametric or non-parametric nature. I cannot suggest what would have been the most appropriate test in the circumstances, since the necessary descriptive statistics are not complete enough to permit confirmation of the normal distribution of data here.*

   The results of the t-test are not presented in any of the conventional ways: there is no evidence of the t values, the s.d. statistic, or the df. The researcher goes on to calculate any improvement as a result of receiving the experimental feedback by subtracting the experimental group means for each error type from those of the control group. This process of calculation, arriving at a negative or positive figure to establish improvement or otherwise, confirms that the success of the software feedback is to be

measured as a consequence only of increasing or decreasing frequency counts, rather than of any variation in the quantity and type of error made. I have made a note here to see whether the researcher subsequently picks up on this limitation on outcomes.

ii. *In the previous appraisal, I made a note to check for more information in this section on the assumptions of normality and equal variances. It is impossible for me to imagine whether a normal distribution has been obtained here in the data, since insufficient information is provided about dispersion of scores. As no s.d. statistic is reported, it is equally impossible to confirm whether variances are equal. However, since group sizes are highly unequal, it does seem as though this latter assumption might not have been met in the study.*

6. What is your appraisal of any other **interpretations that the researcher makes** of his or her data in this section?
In general, the information provided by the tables is not expanded upon in the body of the text. The researcher comments after Table 2 that the number of errors "decreased at a significant rate for both groups the more…time went on". However, there is no evidence for this presented in any table, and no data recorded for any between-letter comparisons. Indeed, this would have been an important statistic to calculate, for it would tell the researcher the extent to which both groups were improving (i.e., with or without the software feedback). I cannot see where the statistic of −1.14 comes from to show that the experimental group decrease is greater than that of the control group. There would appear to be implicit acknowledgement here that the control group were also committing less errors, and I wonder if the unspecified personal teacher feedback they received was helping and/or perhaps their normal L2 classes had an impact on results.

"Spelling" is claimed to be the type most susceptible to improvement with the use of the program. I wondered if we can conclude that the link between software feedback and a reduction in errors is such a direct one. Presumably, a subject writing a totally different letter in a very different genre will be using different lexis to communicate content. Will error incidence and feedback on these in a previous letter necessarily act to lessen the number of spelling errors in the next letter? There might be direct help in the unlikely case that the subject chooses to use the same lexis as in the previous letter and remembers what the program told him or her. Otherwise, any decrease in spelling errors in the next letter might also be due to the writer's heightened awareness about spelling errors as a result of using the program.

Finally, the researcher suggests the reason why there were more "Punctuation" errors registered in the experimental group was because the computer program was not able to supply the kind of knowledge required to see an improvement. I read into this that the help offered by his or her program may be limited to certain error types and not others. If, as it now seems, the program is only partially successful in helping decrease some errors in writing, and if the teacher will still need to input knowledge in other key areas of deviance, I wonder how useful the software program will eventually be in alleviating the kind of problems teachers of L2 writing face in Country X (see Observation 4 below).

7. What information is made available or can be calculated about **effect size** of the outcomes?
   In the light of the above answers, it is impossible formally to calculate effect size or power from the data made available. An informal judgement of this can be made from the actual sizes of differences between means, although data are still needed here to make such an estimate.

8. What initial conclusions do you come to about the **practical significance and meaningfulness** of these results? Do these coincide with the researcher's interpretation?
   Firstly, if the program is only effective to a measurable extent with a restricted number of the 45 error types tested here, I wonder how useful it can really be claimed to be. Furthermore, I feel more confidence might have been placed in the software program (given the positive outcomes mentioned in the literature review from previous research with these kinds of programs) and a higher alpha level predicted. The chances of obtaining a statistically significant difference increases with sample size and statistical significance could have been obtained here at the cost of practical significance.

   I read that final computations of differences for the 45 error types showed that the experimental group had less errors for 32 types; however, I am not shown all these figures, so it becomes impossible to consider how meaningful these differences actually were. Although "Spelling" shows a large difference in favour of the experimental group, the other improvements seem to be very small in comparison, and those I do not see were probably even smaller.

   Finally, the eventual practical merits of a feedback program based on "the most common errors of EFL intermediate level students" would need to be tested in other, less specific, L2 learning situations and in response to other kinds of writing. While this does not weaken the results in this case, I have

made a note to recall this as a perceived limitation on findings in the subsequent section.

Observations

① Comment on the kind of errors possibly contained in the "Typical" category.
The researcher asks the reader to accept the apparent significance of "Typical" errors in the overall ranking and in Tables 3–5 with no information about their content or provenance other than that they were "a special category…based on a data-base of over two hundred errors" from the researcher's own previous marking of compositions. With so many errors apparently falling into this category, and with 44 other types to choose from, I wondered if there was no overlap with other types, and how any such doubtful cases were resolved in terms of grading.

② Think about any overlap between the categories mentioned here and elsewhere.
Once again, it would have been important to see — not least for replication purposes — a more detailed description of error types, for me to understand how apparently similar or inclusive types such as "Adverbs" and "Poor adverbs" or "Punctuation" and "Poor commas" were differentiated in practice.

③ Comment on the implications of "therefore".
This seems to imply there will be a positive direct effect found of using this program on the frequency of the errors mentioned in the previous sentences. This strikes me as a doubtful claim with so much information lacking about the research design and materials. There could be a more subtle indirect interaction going on between letter genre, feedback, and the type of student here (i.e., ESP-oriented).

④ Comment on the implications of "…and so…".
I am not sure that the relative amount of time spent by the teacher correcting spelling will decrease as a result of students using the feedback. Firstly, there is no information available about the comparative times spent correcting exercises, or giving feedback to the experimental and control groups. Secondly, the teacher would still have to correct other items that may take up just as much, if not more, time. The program might help clean up those items that were previously detected as incorrect, but it will not necessarily help prevent further deviance in other words.

III Results: The presentation and nature of findings 2 (Guided appraisal)

1. Before working through this section of the paper, you should re-read the corresponding "RESEARCH QUESTIONS AND HYPOTHESES, VARIABLES, AND OPERATIONAL DEFINITIONS", "SUBJECTS AND MATERIALS", and "PROCEDURES, RESEARCH DESIGN, AND DATA ANALYSIS" texts.

2. Read this section below. As you are reading, in the *right-hand column*, write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face. Advice on how to go about reacting spontaneously to the text was provided on p. 153.

First of all, I compared the two scores obtained on the pre-test measure and the post-test measure for the 84 subjects in the experimental group (classes 1 and 2) and 34 in the control group (class 3). The figures in Table 1 below show these scores, together with the increases or decreases in the means for each group. The maximum score available in the holistic scale was 17. Pre-test results show that the control group had the highest mean (11.14), together with a relatively small variability (s.d. = 1.32) when compared with both experimental groups. The post-test results, however, present a very different picture: now the mean for one of the experimental groups (class 2) is the highest of the three (12.79). By the end of the study period, the improvement for the control group is only .74, but in both experimental groups this improvement is registered at least at over four times this figure.	

Table 1. Pre/Post-test scores and gains for experimental and control groups.

Group	Pre-test mean	s.d.	Post-test mean	s.d.	Points gained	s.d.
Experimental (Class 1)	8.01	1.29	11.26	1.44	3.25	1.31
Experimental (Class 2)	8.79	2.31	12.79	2.40	4	2.98
Control (Class 3)	11.14	1.32	11.88	2.29	0.74	2.86

Data from the two experimental groups show that Class 2 improved considerably more than the other groups. Indeed, they already were higher than the other experimental group in the pre-test. Having said this, this group also showed the greatest variability of scores. Indeed, Class 2 was seen to be a very diverse group before the experiment started. Class 1, on the other hand, would have been less interested in English as they were not going on to further study in this subject. Table 2 shows graphically the improvement/deterioration in each of the experimental groups.

Table 2. Pre/Post-test improvement/deterioration for experimental groups.

Pre-test to Post-test	Class 1	Class 2
−3	4	1
−2	2	0
−1	2	3
0	7	2
1	3	5
2	3	5
3	6	8
4	6	3
5	4	3
6	2	7
7	2	2
8	0	2
9	0	2
	n = 41	n = 43

It is clear from this table that 13 students in Class 2 made an improvement over the two tests of six or more points, compared to only 4 in Class 1.

Questionnaire data back up the gains made by Class 2 in that many subjects in that group made a point of saying that the instruction was enjoyable and motivated them to revise more. Those interviewed who had made the greatest improvement said they felt that using these strategies helped them directly to improve their writing. Having said this, overall questionnaire returns from both experimental groups showed a majority preferring a return to the traditional system while recognising that the revision instruction might be useful in the Cambridge examination. Finally, there was a clear trend in these responses towards understanding writing more in terms of both content *and* accuracy after the instruction was completed.

The few interviews carried out threw more light on some of these questionnaire responses. Most subjects interviewed suggested that the instruction had helped them to account for audience in their writing. They thought it had become easier to see what the teacher wanted in their work since the instruction was felt to have revealed the reasons behind the evaluation that went on ①. Interestingly, several subjects remarked that they saw the new instruction as contrasting with the normal way of focussing on accuracy in evaluation, and three expressed their worries about whether employing the method in the future would act against them once out of the experimental period ②.

3. **Read the textbook introduction to this section again and then respond to these questions, using some of my prompts if you wish:**

1. Does your initial reading of this section suggest that enough data have been provided so as to have **adequately responded to the research questions or hypotheses** previously put forward?
2. What **tables or graphical displays** of results are provided, and what do you understand from the data displayed? Are there any data that you feel might have usefully been added to the information provided here?
   What do you think might have brought about the initial inequality in the groups?

Consider alternative explanations for the improvement/non-improvement shown.

3. What information is provided by any descriptive statistics about the **distribution of data**?
4. Have the data **been scored using the unit measurement predicted** earlier, and/or has any appropriate **data conversion** taken place?
5. i. What, if any, **specific statistical operations or calculations** were carried out on these data, and does this seem to have been carried out appropriately?
   ii. In the light of what you have read in this section, do you wish to amend or add to your previous appraisal of the **assumptions met** for this procedure?
6. What is your appraisal of any other **interpretations that the researcher makes** of his or her data in this section?
   How justified is the concentration on Class 2 for initial interpretations of data? How do you react to the majority comment from both experimental groups that they would prefer to return to the traditional teaching method?
7. What information is made available or can be calculated about **effect size** of the outcomes?
8. What initial conclusions do you come to about the **practical significance and meaningfulness** of these results? Do these coincide with the researcher's interpretation?

Observations

① Comment on the apparent validity of these data in terms of who made the comments, how they were analysed, and when they were obtained.

② In what ways do you think the demands of the present writing context might affect the wider introduction of the experimental treatment?

IV The quality of the discussion and conclusions 1 (Worked sample appraisal)

1. **Before working through this part of the paper, you should re-read the previously-appraised sections of this study.**

2. **Read this section below. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face.**

In this worked example, the columns have been filled in to show you how you might go about doing this initial reading task.

Results support previous findings; software use leads to reduction in errors and better understanding of L2 writing process in specific genres	The findings here support results from our previous studies①, wherein we also saw that feedback from the computer brought about a reduction in the number of errors revealed. Since results here showed significant reductions in the test group's errors compared to the control group, it is clear that using such software and its subsequent feedback can serve as the starting point for a better understanding of how students go about writing and the way they do this in specific writing genres.	*Sorry, I don't see the connection — how does the use of the software, as explained here, help us "understand" more about the L2 writing process?*
Separate error profiles showed few differences between groups, but cumulatively test group were better by the end.	A different profile of subjects' errors was presented after each of the five letters written. There were few differences noted in these profiles between the experimental and control groups, although the test group had significantly fewer error types by the end of the study period ②. Some of the 45 types monitored by the program showed excellent responses in terms of their reduction.	*How much information was shown us in these "profiles"?* *Is there any reason why some showed better responses than others?*

Advantages for the teacher of writing in large classes: information provided which helps predict student errors ⇒ less correction for teacher. Objective is to create more error profiles to help teachers in class preparation.	Pedagogical implications As regards teaching practice, there can only be advantages for any teacher of writing in large-class situations like those in Country X ③. It has become clear that, by analysing the kind of information provided by this software, a teacher can obtain valuable information about where his or her students are most likely to commit errors in their writing in these five kinds of genre; consequently, these teachers will feel less weighed down by the amount of correction they normally would have to do. Ideally, what now needs to happen is for research to go on getting more results from more students in more teaching contexts in this country until we have obtained many thousands of profiles, rather than "only" the 374 obtained here. After many years of working with, and obtaining information from, this program, a teacher could then use these profiles to guide him or her in the preparation of classes. It is clear that certain error types are helped by the use of the program, and the teacher can use specially-designed classes to handle the more resistant errors such as those of punctuation and sentence variety in a more traditional way in the class.	*…but this will not help "any teacher of writing" — as you say later, it depends on the kind of writing/genre being produced.* *How long is "many years"? Are you saying that the eventual benefits of the software are such a long way off?*

3. **Read again the relevant section in the textbook introduction, and then study my responses to the following questions:**

1. What **conclusions** were drawn from the study, and how do these reflect on the **original questions and/or hypotheses**?
   As "…feedback from the computer brought about a reduction in the number of errors revealed", the directional intention towards improvement implicit in the original hypothesis is now said to have been confirmed. While this summary does respond in part to the original statement of intent, I think it might also over-simplify the kind of results obtained. I remember that improvement was noted in both control and experimental groups here, albeit the latter at a faster rate. This somewhat unexpected outcome for both groups has not been addressed here or previously in the paper, despite the fact that it apparently

detracts somewhat from the perceived effect of the software program (see below, question 2). Observations will also need to be made concerning the extent to which these results shed light on the problems envisaged in the "Background to the problem and the problem statement" (see below).

2. What is your appraisal of the **general inferences** which the researcher draws from the findings? How do these **compare with your own reactions** to what you have been told throughout the paper?
   I wondered how justified is the inference that computer feedback "…brought about a reduction…", since it could be understood from this that the software has actually caused the outcomes. Indeed, previous references to this software in the literature review also speak of a program that "produced" (p. 166) overall improvement. However, the research design employed simply does not allow us to say for certain what led to any improvement. Any of the following factors, or some subtle combination of them, could have intervened in the experimental group performance: attitudes towards using the computer, the method of using the software feedback, the amount of time each student interacted with the feedback, the personal teacher feedback obtained, and even the nature of normal class teaching. Also, the understanding I gained in the "Results" section was that improvement has come about basically through the amount of contact with the feedback, rather than the quality of that contact. Finally, I wonder how far the overall success of the program may be being judged as a result of the effect on relatively few error types, rather than across all 45 original categories.

   The researcher also goes on to widen the perspective of his or her interpretation by suggesting that "since" there were significant reductions in error frequency, the software and its accompanying feedback can help us to understand "how students go about writing and the way they do this in specific writing genres". I do not see a logical link between the reason (i.e., "significant reductions in error" as a result of using the feedback) and the claim that we are thereby able to have more insight into the individual writing process. These error frequency counts tell me something about the product of the group writing process on the paper; I do not see what has been learnt here about the way individual writers actually process errors or go about correcting, for example.

3. In what ways are the findings **related to current theoretical and empirical knowledge** on the topic?
   There is no direct re-assessment of the literature in the light of these findings, beyond a claim that the authors' previous findings in Country X have thereby

been supported (see below, Observation 1). I was told that computer-generated feedback had received "positive results" with L1 speakers, and clearly these outcomes will now provide more support from an L2 context. The author also mentioned the need for data from "EFL/ESL students over longer periods of study", but the relevance of this longitudinal study is not subsequently elaborated upon here.

The researcher does not take up his or her own suggestion that these data might contribute to what is already known about computer-assisted language learning (CALL) and L2 writing. Perhaps these findings could have been related to current theory about the process nature of L2 writing. For example, in the process model, error correction can take place at any stage in the process of composition, rather than only once writing has been completed. It might have been interesting to know whether the software and subsequent feedback provided "as they are working" (see p. 180) encouraged the experimental group to proofread for errors only before handing in their finished work or to undertake more subtle revisions while they were actually writing.

4. What **limitations or weaknesses** have you or the researcher identified, and how might any **future research** seek to contribute further to what has been revealed in the study?
*There is no statement here about possible limitations of the results or weaknesses in the study. My reading of the text indicates a continuing desire to generalise outcomes: for example, "it is clear……can serve as the starting point for a better understanding of how students…" or that there are "…advantages for any teacher of writing in large class situations like those in Country X.." (*my underlining*). However, external validity threats were apparently not met in this design (see p. 195), specifically with regard to subject selection and assignment to groups. In such circumstances, the findings could, at most, only be applied to groups of a similar nature and provenance and in a similar context in Country X.*

A number of problems were envisaged with regard to history factors affecting the sample. In particular, it remains unclear to me if (and how) the researcher had been able to ensure that the only L2 writing performed, and the only feedback available on subjects' writing, during the whole six-month period of the study actually came from the specific procedures in the experiment.

Little information has become available about the functioning of the software. I have argued that any hypothesised benefit from using the software might, therefore, be very much dependent on individual interaction with it. There is also little address of the results that showed the control group improving

over the experimental group. It may be of concern, for example, that three of the most-encountered errors across classes and assignments reported in the descriptive statistics (Table 1: "Punctuation", "Sentence variety", and "Typical") ended up registering higher incidence rates in the experimental group than the control group.

The researcher sees the way ahead for further research as "getting more results from more students in more teaching contexts in this country until we have obtained many thousands of profiles…". I would like to have seen some specification of what teaching contexts should ideally be studied, and why. For example, does this mean the researcher sees teaching context as potentially intervening in the perceived link between software use and improvement? I also think this software offers the opportunity for researchers to obtain more specific information on individual writing performance and error correction by concentrating on the individual processing of the feedback, rather than solely on the end product in terms of error frequency.

5. What is your appraisal of any **practical inferences** which the researcher draws from the study in terms of **pedagogical implications or recommendations**?
The practical conclusion drawn is that the software helps the teacher because the analysis of its outcomes provides him or her with "..valuable information about where…students are most likely to commit errors…in these five kinds of genre". The implication is that the usefulness resides in its capacity to predict error occurrence. However, what I have read here are descriptions of these particular students' errors using the software; no evidence has been presented that suggests the software can accurately predict errors. Moreover, although large numbers of subjects have been used here, they are all studying in one specific area, so their results can hardly be said to help other teachers in other situations to predict the errors of their own students.

Interestingly, the researcher goes on to conclude ("…consequently…") that this feature of the software is where teachers themselves will most benefit. As early as the abstract, the researcher talked of an aim to "alleviate teacher obligation in the correction of written work" in large classes. My suggestion was that the instrument might be relatively effective in alleviating part of the problem, but not solving it. What follows here may arguably signal more work for the teacher: apparently "more resistant" errors — upon which the program seemed to have less effect — are recommended to be the subject of more detailed "traditional" attention in class, anyway.

6. Are there any **additional points raised during your appraisal** of the paper that you would like to have seen discussed in this section?
I did wonder whether the program devised by the author was itself bringing about an effect on error frequency or whether this effect was in some way due to the addition of the parsing program Grammatik©. The way the two programs differed and were subsequently integrated has never been clarified for the reader. Since very little information is given about the specially-designed software itself, it remains unclear what the combined effect of these two programs actually was. I also questioned earlier what "feedback" a computer might usefully give, beyond indicating the error and, perhaps, providing the correct version. I wondered whether future research with this instrument might usefully consider studying how far a combination of machine response and directed individual "traditional" teacher feedback could bring about improvement.

Observations

① What more information could have been useful here about these studies?
It is not immediately clear which study is being referred to. In the literature review, I read about work (References 12 and 13) using computer programs, although I did not read that this involved the kind of feedback mentioned here. Indeed, since these are the only references to previous work using computers in L2 writing in Country X, it would have been interesting to read how the present results compared with these studies.

② Comment on the use of "significantly" in the context of this study.
This may be read as indicating that some statistical procedure revealed such differences. However, the "Results" section only talked of "..errors that were less than the control group's for 32 types" (p. 209). I do recognise, however, that this was a majority of types (out of the 45 monitored by the program). Unfortunately, as the "some" in the next sentence reveals, the feedback can also be said to produce different degrees of response depending on the error.

③ Has enough information about operational definitions been provided in the paper to support this?
At the beginning of my appraisal of this paper, I noted that very little was revealed about what actually constitutes "large-class situations like those in Country X". Clearly, massification in classes is highlighted throughout, and the "Background to the problem and the problem statement" talked of the pressures of getting students to examination levels in L2 proficiency. However, how

this translates into actual class numbers and the specific teaching and correcting responsibilities of the instructor is unknown. Consequently, at this point, it is impossible for an interested reader to compare his or her own research context with that of the study and decide whether the outcomes reported are, indeed, likely to be as advantageous.

IV The quality of the discussion and conclusions 2 (Guided appraisal)

1. **Before working through this part of the paper, you should re-read the previously-appraised sections of this study.**

2. **Read this section below. When you have read a paragraph, stop and write in the *left-hand column* a few words which summarise the gist of that paragraph, to help you understand and focus on what the researcher is saying. Then, in the *right-hand column* write a few words which record your instinctive reactions to what you have just read as if you were talking to the researcher face-to-face. Advice on how to go about reacting spontaneously to the text was provided on p. 153.**

	The results of the study were as expected: the experimental groups improved to a greater extent in the post-test measurement, even though the control group scored highly on the pre-test. Results from the questionnaire data showed that subjects were being attentive to a wider range of factors ① in their writing as a result of the revision strategy teaching. Having said this, data from the interviews tended to contradict this; some subjects expressed a narrower view of revision which reflected the kind of instruction they received in their normal L2 writing classes.	

What was clear, however, from both questionnaire and interview data was that subjects had begun to form an awareness of the reader. Admittedly, they may still have seen this reader in terms of their immediate writing context, in other words, their teacher. As Researcher 13 (1995) had found, subjects showed a certain anxiety about writing grammatically correct forms; the interview data reveals subjects who expressed this anxiety in terms of the possible repercussions on any eventual exam grade. The impression gained was that they were still unsure about whether this kind of revision would fit in with the way their compositions were normally evaluated.

Pedagogical implications

These findings support the view that the teaching of L2 writing needs to take into account a number of factors, and not just surface accuracy. The common practice (Researcher 20, 1981; Researcher 21, 1983) ② of encouraging initial L2 writing training by concentrating on surface correctness needs to be amended to include more instruction about discourse-related skills and/or how to focus on content in revision. Improvement in the control group was not as much as the experimental group; therefore, it seems that writing more, and on more subjects, and correcting surface errors only does not actually help to improve writing.

	Limitations Although findings do indeed suggest that the experimental groups profited from the instruction in terms of their points gain, it must also be admitted that no control was made in the study for what happened in each separate lesson or for the gains which might naturally accrue as a result of their normal English learning during the period ③. Also, it should be recalled that only a few students were interviewed and that these were volunteers. Thus, these cannot be said to represent the views of the whole sample. Since this study only sought to describe what happened in one experimental situation, more research is needed to establish firm tendencies. Encouraging more reader-based views about revision means encouraging the kind of peer reading and writing integration built in to the experimental teaching given here. Thus, as the experimental groups were reading each other's texts, they were also being taught revision strategies. The control group did not receive such benefits, and this could explain the comparatively small gains made overall pre- to post-test. It follows from this evidence that revision instruction certainly seems to be a recommendable practice. These groups of subjects were able to stand back from their work and assess the need for more global revision as a result of this instruction. The question for curriculum writers in Country B is whether an alternative should now be sought to the current practice of only encouraging the kind of frequent L2 writing in class practised here by the control group.	

3. **Read the textbook introduction to this section again and then respond to these questions, using some of my prompts if you wish:**

1. What **conclusions** were drawn from the study, and how do these reflect on the **original questions and/or hypotheses**?

2. What is your appraisal of the **general inferences** which the researcher draws from the findings? How do these **compare with your own reactions** to what you have been told throughout the paper?
   Consider again the implied link between instruction and improvement here?
   What other factors might have contributed to subjects' perceptions about the writing and revision process?
   What other factors might have contributed to improvement in the main study?
   Are you satisfied with the conclusions drawn about the control group's performance?
3. In what ways are the findings **related to current theoretical and empirical knowledge** on the topic?
4. What **limitations or weaknesses** have you or the researcher identified, and how might any **future research** seek to contribute further to what has been revealed in the study?
   How might the composition of the final group of interviewees have affected responses obtained?
   Consider any limitations imposed on the study by the school authorities.
5. What is your appraisal of any **practical inferences** which the researcher draws from the study in terms of **pedagogical implications or recommendations**?
   How appropriate are the recommendations made in the light of current constraints on writing in Country B?
6. Are there any **additional points raised during your appraisal** of the paper that you would like to have seen discussed in this section?
   What more details would have been useful about the revision strategies taught?

Observations

① What information has been provided about these "factors" in the paper?

② What might be deduced from the dates of the research cited here compared to when the paper was written (1998)?

③ To which groups does this limitation refer?

Glossary of key terms in quantitative research

Alpha (decision) level (α) – (see "Statistical significance level")

ANCOVA (see ANOVA)

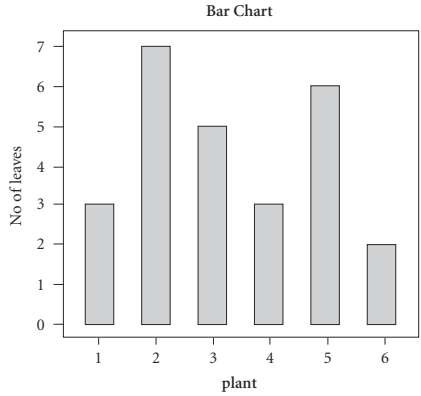
ANOVA – Analysis of variance (ANOVA) is used to test hypotheses about the differences between two or more means. The t-test can only be used to test differences between two means. When there are more than two, it is possible to compare each mean with each other mean using t-tests. However, conducting multiple t-tests can easily increase the possibility of making a Type I error (see below). Analysis of variance can be used to test differences among several means for significance without increasing the chances of committing a Type I error.

Among the statistical procedures used in the analysis of data are the analysis of variance (ANOVA), analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA), and multivariate analysis of covariance (MANCOVA). These procedures aim to divide the total variance into its components by the analysis of sums of squared terms taken from the data.

Assumptions – Statistical methods such as those we have studied in this book require the data to satisfy various conditions, for example that the data follow a normal distribution and are independent. When using such a method, we assume that these conditions hold: these are the assumptions required for the method to be valid. It is good statistical practice to check the assumptions as far as possible.

Balanced design – Another important aspect of the two-way ANOVA test is related to the number of subjects in each cell of the design. If the between-groups independent variable has equal numbers of subjects in each of its levels, then you have a balanced design. With a balanced design, each of the two independent variables and the interaction are independent of each other. Each independent variable can be significant or non-significant, and the interaction can be significant or non-significant without any influence from one or the other effects.

Bar graph (chart) – A bar graph is similar to a histogram, except that there is a small space drawn between the columns. It is often used in summarising a set of categorical data. A number of rectangles are used, all of the same width, each of which represents a particular category. The length (and hence area) of each rectangle is proportional to the number of cases in the category it represents.



Bell-shaped curve (see "Normal distribution")

Between-groups designs – Between-group variables are independent variables or factors in which a different group of subjects is used for each level of the variable. If an experiment is carried out comparing four teaching methods and if a different group of subjects is used for each, then teaching method is a between-groups variable. If every variable in an experimental design is a between-groups variable, then the design is called a between-groups design. Some experimental designs have both between- and within-group variables (cf., "Mixed designs"; "Within-group designs").

Bi-modal distribution – A bi-modal distribution has two modes.

Category (variable) – A set of data is said to be categorical if the values or observations belonging to it can be sorted according to category. Every value should belong to one and only one category, and there should be no doubt as to which one. For example, people have the characteristic of 'gender' with categories 'male' and 'female' (see also "Nominal data measurement").

Causal relationships – If there has been an identified probability of one event influencing another event, it is suggested that a causal relationship exists between the two events. In order to attribute a causal relationship between two events, A and B, three conditions are typical: (a) B must not precede A in time, (b) A and B must covary together to a recognisable degree, and (c) no alternative explanation accounts as well as or better for the covariation between A and B.

Central tendency – Measures of central tendency locate the middle or the centre of a distribution of data. There is deliberate vagueness about the way "middle" or "centre" are defined. Thus, the term "central tendency" can refer to a wide variety of measures. The mean is the most commonly used measure of central tendency. Others are the "median" and the "mode". For normal distributions, these measures are all the same. For skewed distributions, they can differ considerably (see also "Mean"; "Median"; "Mode")

Chi-square(d) test – The chi-squared (or chi-square) test of independence is a test of whether there is a relationship, for example, between subjects' characteristics on one variable and those on another. The test is based on the chi-squared distribution, the most common use of which is to test differences between proportions. Although this test is not the only one based on the chi- squared distribution, it has come to be known as the chi-squared test. The test compares the observed frequencies in each of the cells of a contingency table with the expected frequencies for each cell if these differences were only due to chance. The greater the difference between observed and expected frequencies, the more likely the result is to be significant.

Confounded research design – Two variables are confounded if they vary together in such a way that it is impossible to work out which one is responsible for an observed effect. For example, imagine a study wherein two L2 teaching methodologies were compared. The first was given to a group of teenage students and the second to a group of adults. If a difference between treatments were revealed, it would be impossible to tell if one treatment were more effective than the other, or if teaching methodology treatments are more effective for one age group than the other. In such an example, age and treatment would have been confounded.

Construct – An abstract theoretical concept that is not directly observable or measurable (e.g., motivation, language-learning aptitude) but that is considered to exist on theoretical grounds.

Construct validity – Construct validity describes

the extent to which a particular instrument measures accurately constructs of interest that have been obtained theoretically (see also "Content validity"; "Face validity"; "External validity"; "Internal validity")

Content validity – Content validity considers formally the extent to which a particular instrument measures accurately what it is claimed to measure. A group of experts would normally decide on this, focussing on the instrument's representativeness and comprehensiveness (see also "Construct validity"; "Face validity"; "External validity"; "Internal validity") .

Continuous data measurement – Continuous data show us how much of a variable is present in the set of data. It would be possible to score any value, within the limits to which the variable extends. You can count, order, and measure continuous data. For example, the variable "Amount of time needed to complete a test of reading comprehension" is a continuous data measurement/variable as it could take 30 minutes, 35 minutes, etc. to finish. There is no set time limit. However, the variable "Number of correct responses on a reading comprehension test (Total score possible 50 points)" would not be a continuous data measurement/variable, as it would not be possible to get 32.15 on such a test. A variable that is not continuous is also called "discrete".

Control group – Sometimes termed the "Comparison" group in a quasi-experimental study with no random assignment to groups, this refers to the group in a quasi- or pure experimental study that does not receive the treatment, later to be compared to the experimental or treatment group. In a pure experimental study, subjects are allocated randomly to the treatment and control groups.

Control variable – The effects of a particular variable may be isolated by controlling its presence or its consequences. This is done by controlling its potential effect on the dependent variable. However, controlling in this way also places inevitable limits on generalisation of outcomes, since the researcher will not be able to generalise beyond the controlled situation in the study (see also "Independent variable"; "Dependent variable"; "Moderator variable"; "Intervening variable").

Correlation – The correlation between two variables represents the degree to which variables are related. Typically, the linear relationship is measured with either Pearson's correlation or Spearman's rho. It is important to keep in mind that correlation does not necessarily mean causation. For example, there may be a high positive relationship between the number of ambulances attending a major car accident and the number of people injured. Does this therefore mean that the ambulances cause the injured? It is more probable that the larger the accident, the more ambulances attend. Thus, the variable "seriousness of car accident" is the causal variable, correlating with the number of ambulances attending the scene (see also "Negative correlation"; "Pearson correlation coefficient"; "Spearman correlation coefficient"; "Linearity"; "Scatterplot") .

Correlation coefficient – A correlation coefficient is a number between −1 and 1 measuring the extent to which two variables have a linear relationship. A correlation coefficient of 1 is obtained if there is perfect linear relationship with a positive slope between the two variables. In the case of a positive correlation, whenever one variable has a high (or low) value, so does the other. A coefficient of −1 is obtained if there is a perfect linear relationship with negative slope between the two variables. In this case, whenever one variable has a high (or low) value, the other has a low (or high) value. There are a number of different correlation coefficients appropriate to the different kinds of variables being studied (see "Correlation"; "Negative correlation").

Covariate – A variable can be applied in an

analysis to correct, adjust, or modify the data on a dependent variable before these data are related to one or more independent variables. For example, if a researcher was looking at the relationship between student age and L2 examination success, the researcher might first want to remove any effects due to the amount of time spent studying the L2. This latter would then be the covariate used.

Crit(ical) values – A critical value is used in significance testing. In most procedures, it is the value that a test statistic must exceed in order for the null hypothesis to be rejected. For example, if you look at the statistical tables for the t-test distribution, the critical value of t (with 5 degrees of freedom using the .01 significance level (two-tailed hypothesis)) is 4.03. This means that for the probability value to be less than or equal to .01, the value of the t statistic obtained from the calculation must be 4.03 or greater. The critical value for any hypothesis test depends on the significance level at which the test is carried out, and whether the test is one- or two-tailed.

Curvilinear – (see "Linearity")

Data – Information collected in a research study is referred to as data. Data are often considered to be statistical or quantitative. Data can, however, also be found in many other forms, including transcripts of interviews and videotapes. Different kinds of data require different approaches to statistical analysis (see "Nominal data measurement", "Ordinal data measurement", "Interval data measurement").

Degrees of freedom (df) – This value is frequently referred to in the organisation of tables of statistical distributions used in carrying out tests of statistical significance.

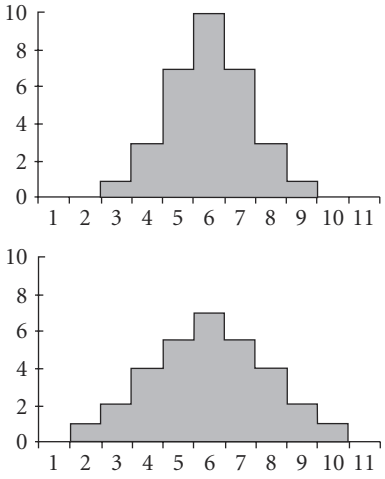
Dependent variable – A variable in a study, whose values are "dependent on" other variables for their outcomes. The researcher would try to explain these outcomes in terms of one or more independent variables. The distinction between dependent and independent variables is typically made on theoretical grounds to test a particular model of cause-effect or a specific hypothesis. You may also come across the term "criterion variable" to describe this variable (see also "Independent variable"; "Control variable"; "Moderator variable"; "Intervening variable").

Descriptive statistics – A basic use of descriptive statistics is to summarise a mass of data in a clear and understandable way, both numerically and graphically. In numerical presentations, we might typically look out for measures of central tendency and variability, such as the mean and standard deviation. These statistics convey information about the most typical values obtained and how these are spread out across the data sample (see also "Inferential statistics").

Directional hypothesis – When the researcher wishes to assess or compute the probability of differences in both directions of the distribution, this is referred to as a "two-tailed" hypothesis. However, other situations (for example, as a result of previous studies with these variables) may suggest to the researcher that he or she need only look in one direction for the probability of these differences. When only one direction is of concern to the researcher, a "one-tailed" test can be performed. When consulting statistical tables yourself, be sure to confirm whether these correspond to one-tailed or two-tailed hypotheses (see also "Hypothesis testing").

Dispersion – The dispersion (variability or spread) of scores from a variable is the degree to which these differ from each other. If every score on the variable were about equal, there would be very little dispersion noted. When the dispersion is large, the values are more widely scattered (bottom diagram); when it is small, they are more bunched around one point (top diagram). There are several measures of dispersion, two of the most common being the "stan-

dard deviation" and "range" (see below).



Effect size – An effect size is a standardised measure of the strength of a relationship. Its great advantage to the researcher is that the measure is independent of sample size and estimates the extent to which the phenomenon takes place. A number of different measures are used to test the effect size. The larger the effect size, the easier it is to detect (and, therefore, the fewer cases needed to do so). See also "Power".

Eta² – Eta² is a correlation coefficient that can be used to determine strength of association or effect size. It is interpreted as the proportion of the total variability of the dependent variable which is explained by the variation in the independent variable (cf., "Omega²").

Expected frequencies – In contingency tables (as presented in a chi-squared test), the expected frequencies are the frequencies that you would predict ('expect') in each cell of the table, if you knew only the row and column totals, and if everything were equal as it would be if there were no relationship at all between the variables (see also "Chi-squared-test"; "Observed frequencies").

Ex post facto designs – A design in which the researcher — rather than creating the treatment to be tested — examines the effect of a naturally-occurring "treatment" after that treatment should have taken place (i.e., "after-the-fact"). This "treatment" is then related to some result or dependent measure. If a predicted relationship is confirmed statistically, this will not necessarily be an indication that the independent and dependent variables are causally related (cf., "Pure experimental designs"; "Pre-experimental designs"; "Quasi-experimental designs").

External validity – A study would have external validity if the findings could be applied in the real world (i.e., outside the current experimental situation) and to similar events as in the present study. The extent to which threats to such validity are met affects our ability to credit the results with generalisable outcomes. External validity is of little value unless it has been proceeded by adequate address of internal validity concerns, which give us confidence in the basic descriptive conclusions drawn from the data themselves (see also "Content validity"; "Face validity"; "Construct validity"; "Internal validity") .

Face validity – Face validity relates to content validity, but assesses informally and/or intuitively whether the instrument appears to measure what it purports to measure (see also "Content validity"; "Construct validity"; "External validity"; "Internal validity").

.

Factorial ANOVA – Two Way Analysis of Variance (ANOVA) is a way of studying the effects of two factors separately (their main effects) and, sometimes, together (their interaction effect). A factorial ANOVA can be designed with many different factors, but by adding more and more independent variables, the potentially numerous interactions can make interpretation difficult.

Factorial design – When a researcher wants to

study the effects of two or more independent variables simultaneously, it makes more sense to manipulate these variables in one experiment than to run a single experiment for each one (such as a t-test). The treatments are combinations of levels of the factors. Moreover, such experiments involving more than one independent variable allow the researcher to test for interactions among variables.

F Distribution – The F distribution is the distribution of the ratio of two estimates of variance and is used to compute probability values in the analysis of variance. The F distribution has two parameters: degrees of freedom numerator and degrees of freedom denominator. In the tables provided, the vertical df corresponds to the within-group measure and the horizontal line across the top to the between-group measure.

F ratio – The F-ratio is the ratio of the between-group variance to the within-group variance in an analysis of variance.

Frequency (data) – Frequency measures how often something occurs, or tallies how many objects, people, or subjects have a particular attribute.

Frequency distribution – Frequency distributions are portrayed in tables, histograms, or polygons. They can show either the actual number of observations falling in each range or the percentage of observations. In the latter instance, the distribution is called a relative frequency distribution (see also "Bar graph (chart)"; "Histogram").

Generalisation – Generalisation refers to the extent to which conclusions can be drawn about a parameter or relationship in a population from data obtained from a sample of that population. Generalisation is biased or constrained when the population from which the sample is drawn is narrow.

Hawthorne effect – The Hawthorne effect refers to the tendency of subjects to improve their performance under observation, simply because they are aware that they are being studied or are involved in an experiment.

Homoscedasticity – This assumption means that the error variance around the regression line is the same for all values of the predictor variable. In multivariate analysis, it is undesirable for the criterion/dependent variable to have variances which are considerably different for the same values of the predictor variable, in the different populations which have been sampled. The incidence of markedly different variances in the different populations is referred to as heteroscedasticity.

Histogram – A histogram is constructed from a frequency table (cf. "Bar graph (chart)"). The intervals are shown on the X-axis and the number of scores in each interval is represented by the height of a rectangle located above that interval. It is generally used when dealing with large data sets. A histogram can also help detect any unusual observations (see "Outlier"), or any gaps in the data.

Hypothesis – A hypothesis is a statement about the relationship between two or more variables that are being studied (see also "Directional hypothesis").

Hypothesis testing – Hypothesis testing consists of estimating the probability that certain, hypothesised effects are observed and calculating these against the null and the alternative hypotheses (see "Null hypothesis"). Data obtained are compared with theoretical expected data, and a calculation made of the probability that the observed outcome could have been due to chance. If the data are very different from what would be expected under the assumption that the null hypothesis is true, then the null hypothesis is rejected. If the data are not so different from those expected under the assumption that the null hypothesis is true, then the null hypothesis is not rejected. If the

researcher does not reject the null hypothesis, this does not mean the null hypothesis is true; it only suggests there is not enough evidence against the null hypothesis in favour of the alternative hypothesis. Rejecting the null hypothesis suggests that the alternative hypothesis **may** be true (see also "Research question").

Independent variable – An independent variable is one that can be used to predict or explain another variable, usually referred to as a dependent (or criterion) variable (see also "Dependent variable"; "Control variable"; "Moderator variable"; "Intervening variable").

Inferential statistics – Inferential statistics are used to draw inferences about a population from a sample (cf., "Descriptive statistics").

Intact groups/classes – Much of the research in L2 learning involves the use of groups or classes into which subjects have been previously placed according to some criterion. Such groups or classes are referred to as "intact". In such research it is impossible randomly to select subjects at the outset (cf., "Random selection").

Interaction (effect) – An interaction occurs when two or more predictor variables not only have separate direct effects, but also a combined effect formed by the product of the two or more variables, which influences the dependent variable.

Internal validity – Internal validity is the extent to which the results of the study can be put down to the treatment applied rather than to the design of the study. It also reflects on the degree to which sound conclusions can be drawn about the results of the study (see also "Content validity"; "Face validity"; "Construct validity"; "External validity").

Inter-rater reliability – The inter-rater reliability of an instrument measures the degree of agreement between two or more raters, and indicates the extent to which the raters assess by using the instrument in the same way (see also "Reliability"; "Kuder-Richardson formulae")).

Interval data measurement – On an interval scale the distance between any two positions is of known size. One unit on the scale represents the same magnitude on the trait or characteristic being measured across the whole range of that scale (cf., "Ordinal data measurement" and "Nominal data measurement"). For example, if language proficiency is measured on an interval scale, a difference between a score of 50 and 51 would be considered to be the same difference as that between 23 and 24.

Intervening variable – The intervening variable is thought to be a predictor of one or more dependent variables. It is a factor that theoretically affects these variables but cannot be seen, measured, or manipulated. Therefore, its effect has to be interpreted from the effects of the independent and moderator variables on the observed phenomenon. Unlike the moderator variable (see below), the intervening variable could not previously have been identified precisely for inclusion in the research (see also "Independent variable"; "Dependent variable"; "Control variable"; "Moderator variable").

Kuder-Richardson formulae – These formulae are measures of the internal consistency or reliability of tests that have dichotomous response categories (e.g. "yes/no", "right/wrong" items). See also "Inter-rater reliability" and "Split-half reliability".

Kurtosis – Kurtosis indicates the extent to which a distribution is more peaked or flat-topped than a normal distribution. The index of kurtosis measures the extent to which the distribution differs from the normal or "bell-shaped" curve (see also "Normal distribution").

Levels (of a variable) – The number of levels of a factor or independent variable is equal to the number of variations of that factor that were used in the experiment. If a researcher were

interested in studying the variable "Native language" and obtained data from "L1 German", "L1 French" and "L1 Russian" students, there would be three levels of the variable.

Likert scale – A Likert type scale is a widely-used questionnaire format in which a series of statements is used to measure a particular characteristic by asking for a response. This usually involves choosing from among two or more possible cues, typically intended to gauge this response, perhaps in terms of the agreement or disagreement with each statement.

Linearity – When two variables are perfectly linearly related, the points of a scatterplot fall on a straight line. The more the points tend to fall along a straight line, the stronger the linear relationship. **Curvilinearity** might occur when the points plotted form a curve rather than a straight line (e.g., the correlation begins highly positive but finishes highly negative). The Pearson correlation coefficient is not suitable when the relationship is curvilinear (see also "Correlation"; "Scatterplot").

Linear regression – Linear regression involves the prediction of one variable from another variable when the relationship between the variables is assumed to be linear (cf., "Multiple regression").

Longitudinal studies – A longitudinal study involves the investigation over time of individuals or groups of individuals. A cross-sectional study is conducted at one single point in time. Most often, the study is of a sample drawn from a population at a particular time, but it may involve the investigation of the total population.

Main effect – The main effect of an independent variable is the effect of the variable alone averaged across the levels of other variables in the experiment. Analysis of variance provides a significance test for the main effect of each variable in the design (cf., "Interaction (effect)").

Matched subjects (group) – Two (or more) samples selected in such a way that each case (e.g., person) in one sample is matched on one or more pre-selected characteristics with a corresponding case in the other sample. Examples often found include two samples in which the members are clearly paired, or are matched explicitly by the researcher, or samples in which the same attribute, or variable, is measured twice on each subject, under different circumstances, commonly referred to as "Repeated-measures" designs.

Mean – The arithmetic mean is what is often referred to as the "average". The mean is the sum of all the scores divided by the number of scores (see also "Central tendency"; Median" ; "Mode").

Median – Once placed in order, the median is the value halfway through the set of data, below and above which we find an equal number of data values. The median is less sensitive to extreme scores than the mean and this makes it a better measure than the mean for highly skewed distributions (see also "Central tendency"; "Mean"; "Mode").

Mixed (group) designs – Also known as "split-plot" designs, these include both comparisons of independent groups (between-groups) and repeated-measures (within-group) of the same group of subjects.

Mode – The mode is the most frequently occurring score in a distribution and is another measure of central tendency. It is the only measure of central tendency that can be used with nominal data. A disadvantage of this measure is that there can be more than one mode if two or more values are equally common (see also "Central tendency"; "Mean"; "Median").

Moderator variable – A moderator variable is a variable in a cause-effect situation that interacts with a prior variable to modify its effect on a dependent variable (see also "Independent

variable"; "Dependent variable"; "Control variable"; "Intervening variable").

Multicolinearity – This occurs when two or more predicting variables are highly correlated in multivariate analyses, and the joint outcome is that they prevent the accurate estimation of the effects of the variables.

Multiple regression – The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. It is often used when the researcher wants to see the degree to which scores on a dependent variable can be predicted from those of two or more independent variables (cf., "Linear regression").

Multivariate analysis – This is a blanket term referring to the family of procedures that involve the simultaneous study of two or more dependent variables in a study. One advantage of such analyses is that they allow the researcher to manipulate the variables and check if any specific group affects the dependent variable more than another.

Negative correlation – A relation in which the values of one variable increase as the values of the other variable decrease (see also "Correlation"; "Correlation coefficient").

Nominal data measurement – Nominal measurement allocates items to groups or categories. Any numbers allocated are merely labels; there is no quantitative information provided. While you can count such outcomes, no ordering of the items is implied. "University degree studied", "place of residence", and "sex" would all be examples of nominal scales (cf., "Ordinal data measurement" and "Interval data measurement").

Non-continuous measurement – (see "Continuous data measurement")

Non-parametric tests – Non-parametric tests are often used in place of their parametric corresponding procedures when certain key assumptions about the underlying population are in doubt. In the case of a comparison, for example, between two independent groups, the Mann-Whitney U test does not have the basic assumptions that the data are strongly interval-based, or that the mean is the best measure of central tendency; its parametric alternative, the independent t-test does (cf., "Parametric tests").

Normal distribution – The normal distribution is a theoretical concept and suggests a particular form for the distribution of the variable which, when plotted on a graph, produces a bell-shaped curve (see below and "Skew(ed) distribution"), rising smoothly from a small number of results at both extremes (the tails) to a large number of cases in the middle. The distribution has certain useful characteristics which lead to its widespread use in statistical tests. Most of these tests work well even if the distribution deviates slightly from normality:

Null hypothesis – The null hypothesis, often written Ho, represents a theory or hunch the researcher has, either because it is believed to be true and/or because it is to be used as a basis for argument, but has not been tested. The null hypothesis suggests an effect does not differ significantly from zero or another set value. The alternative hypothesis (often written H1) postulates that an effect differs significantly from zero or another set value. Depending on the data submitted to the hypothesis test (see "Hypothesis testing"), the null hypothesis either will or will not be rejected as acceptable. The way in which the null hypothesis is communicated is often the opposite of what the researcher actually expects; it is postulated to allow the data to contradict it (e.g., "There is no difference between the test scores obtained

from the experimental and the control groups"). See also "Type 1 and Type 2 errors".

Observed frequencies – In contingency table problems, the observed frequencies are those actually obtained in each cell of the table. Observed frequencies are compared with the expected frequencies and any significant differences between them used to suggest that the example expressed by the expected frequencies does not describe the data well (see also "Chi-squared-test; "Expected frequencies").

Omega² – A strength of association/effect size measure applied in certain statistical tests where the design is balanced (cf., "Eta²").

One-way ANOVA – The one-way analysis of variance allows for comparison of several groups of observations, all of which are independent but possibly with a different mean for each group. In a one-way design, there is only ONE dependent variable and ONE independent variable with three or more levels. The comparisons (within-groups or repeated measures) of the means on the dependent variable are made across these levels (see also "ANOVA").

Operational definition – An unambiguous definition based on the observable characteristics of what is being defined. The precision of this definition affects the nature and quality of the observations upon which the operational definition is based, as well as how they are obtained and subsequently measured.

Ordinal data measurement – Measurements with ordinal scales are ordered in the sense that higher numbers represent higher values. You can count and rank these data, but not measure them. It is important to remember that the intervals between neighbouring points on the scale are not necessarily equal (cf., "Interval data measurement" and "Nominal data measurement"). For example, on a five-point rating scale measuring motivation to learn L2 German, the difference between a rating of 2 and that of 3 may not represent the same difference as that between a rating of 4 and a rating of 5.

Outlier – An outlier is an observation in a set of results which is far removed in value from the others in the data — an unusually large or small value when compared to these. The existence of such cases can have important repercussions on certain statistical tests and distort the interpretation of the data. If an outlier is a genuine result, it is important because it might indicate an extreme of behaviour in the process under study. For this reason, outliers are best examined carefully before embarking on any formal analysis. Outliers should not simply be removed without further explanation or justification.

Parametric tests – A group of statistical techniques that — unlike non-parametric tests — make strong assumptions about the distribution of data from the dependent variable (e.g., that they are normally distributed). Strictly speaking, these tests assume that dependent variables are scored with interval data or data which are strongly continuous (cf., "Non-parametric tests").

Parameters – A parameter is a numerical quantity measuring some aspect of a population of scores. When researchers make calculations to describe a sample, these are called statistics. If these same calculations were made for the population of interest in the study, these would be referred to as parameters. Parameters are rarely known and are usually estimated by statistics computed in samples. Greek letters often symbolise such parameters (e.g., μ = the population mean).

Pearson correlation coefficient (r) – When computed in a sample, it is designated by the letter "r" and is sometimes called "Pearson's r". Pearson's correlation reflects the degree of linear relationship between two variables that have been measured on interval or ratio scales. The resulting coefficient ranges from +1 to −1.

A correlation of +1 means there is a perfect positive linear relationship between variables. However, *r* can be misleadingly small when there is a relationship between the variables, but it is a non-linear one. There are procedures, based on *r*, for making inferences about the correlation coefficient. However, these make the implicit assumption that the two variables are jointly normally distributed. When this assumption is not met, a non-parametric measure such as the Spearman correlation coefficient (see below) might be more suitable (see also "Correlation") .

Phi correlation coefficient (Φ) – Phi is a correlation coefficient calculated between two dichotomous nominal variables.

Point biserial correlation (r_{pbi}) – A point biserial correlation is a correlation coefficient calculated between a dichotomous nominal variable and a continuous (interval) variable.

Polygon – This refers to the visual shape of the distribution of a set of data shown by the (curved) line connecting the plotted points.

Population – A population consists of an entire set of objects, observations, or scores that have something in common. It is the entire group the researcher is interested in, which he or she wishes to describe or draw conclusions about. For example, a population might be defined as all L2 language learners between the ages of 15 and 18. It is important that the investigator carefully and completely defines the population before collecting the sample, including a description of the members to be included. In many cases, such a population is hypothetical. If a researcher in France were testing a new method of learning L2 English vocabulary, for example, he or she might define a population that would be obtained if all teenage L2 learners in France received this new method. Such a population does not exist; the population suggested consists of the scores that would be obtained if they **were** taught with this method (see "Parameters").

Post-test/Pre-test – These do not need to be "tests" as such before and/or after a particular intervention or treatment. These can also take the form of some kind of observation or measurement of the dependent variable.

Power – The power of a statistical test is its ability to detect a significant relationship with a specified number of cases. In other words, the power of a statistical hypothesis test measures that test's ability to reject the null hypothesis when it is actually false — that is, to make a correct decision. It is important to consider power in the design of a quasi-experimental or true experimental study. If the power of an experiment is low, there is then a good chance that the experiment will be inconclusive. There are methods for estimating the power of an experiment before the experiment is conducted. If the power **is** too low, the experiment can be redesigned by changing one of the factors that determine power (see also "Effect size").

Pre-experimental designs – these designs are simple and inexpensive to implement and exploratory in nature, but lack control groups to compare with the experimental group. They are often used in preliminary research to provide direction and focus for further research using experimental designs, or when circumstances exclude more controlled research design. (See also "Ex post facto designs"; "Pure (true) experimental designs"; "Quasi-experimental designs")

Predictor variable – A predictor variable is one used only in correlational relationships for prediction. It can be likened to the independent variable in experimental research.

Probability – A probability provides a quantitative description of the likely occurrence of a particular event. In research, when an hypothesis is offered for testing, an educated guess is being made about what is or is not probable.

An hypothesis is tested by finding out the probability of the result. Probability, once calculated, is the proportion of times a particular outcome would happen if the research were repeated ad infinitum.

Probability value/level (p) – The probability value (*p*) of a statistical hypothesis test is the probability of wrongly rejecting the null hypothesis if it is, in fact, true. In the majority of tests, the p-value is compared with the actual significance level of our test and, if it is smaller, the result is significant. That is, if the null hypothesis were to be rejected at the 0.05 significance level, this would be reported as "$p < 0.05$". Small p-values suggest that the null hypothesis is unlikely to be true. The smaller it is, the more convincing is the rejection of the null hypothesis (see also "Statistical significance").

Pure (True) experimental designs – Criteria for a true experiment are (a) that subjects should be randomly allocated to treatment and control groups, (b) that the treatment and control conditions should be randomly assigned to the groups so formed; and (c) the subjects should be drawn randomly from an identifiable population. The reason for using random selection or random allocation is to exercise control over the identified and unidentified elements that might influence the outcome of the experiment, and to ensure that the results of the experiment can be generalised to an identifiable population. (See also "Ex post facto designs"; "Pre- experimental designs"; Quasi-experimental designs")

Quasi-experimental designs – In quasi-experimental designs, both control and experimental groups are used in the study, but subjects have not normally been randomly selected nor randomly assigned to these groups (see also "Ex post facto designs"; "Pure (true) experimental designs"; "Pre-experimental designs")

Random selection – In random sampling, each item or element of the population is chosen entirely by chance and has an equal chance of being chosen during the selection process. By using random sampling, the likelihood of bias is reduced (cf., "Intact groups/classes").

Range – The range of a sample (or a set of data) is a measure of the spread or the dispersion of the observations. It is calculated from the difference between the largest and the smallest observed value of some quantitative characteristic. However, a great deal of information is ignored when computing the range, since only the largest and the smallest data values are considered. It also follows that the range will be greatly influenced by the presence of just one unusually large or small value in the sample (see "Outlier"; "Standard deviation"; "Dispersion"; "Variance").

Rank order – (see "Ordinal data measurement")

Rate – This is calculated to show how often something occurs in large sets of data. No standard unit is used, as this depends on the nature of the data; however, many studies provide rates such as "per 100 occurrences", "per 1000 words used", etc.

Ratio – Ratio scales are like interval scales except they have true zero values, that is, a point on the scale that represents the complete absence of the characteristic measured.

Raw scores – Raw scores are data or values that have not been subjected to statistical manipulation. They may or may not have been adjusted by having any abnormal values removed or corrected.

Regression line – A regression line is an imaginary line drawn through a scatterplot of two variables, around which most of the plotted points cluster. When it slopes down (from top left to bottom right), there is evidence for a negative or inverse relationship between the variables; when it slopes up, a positive or direct

relationship is indicated (see also "Scatterplot").

Reliability (coefficient) – A measure of how consistent repeated measurements are when performed under comparable conditions. The reliability of an instrument can be discovered in a number of ways and is an index of the consistency or stability with which the instrument makes measurements. Before drawing any conclusions from an experiment, the reliability of the test instruments used in the experiment should have been assessed. However, such reliability can only be judged by administration of the instrument to a sample of subjects. Thus, the coefficient is dependent on the characteristics of the sample, as well as the characteristics of the instrument (see also "Inter-rater reliability"; "Kuder-Richardson formulae").

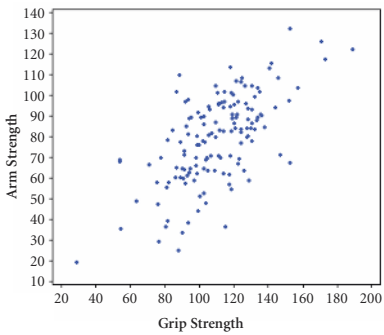
Repeated measures designs – (see "Within-group designs")

Research question – A research question is a specific question asked in the course of investigation to which a specific answer or set of answers is sought (see also "Hypothesis testing").

Sample – A sample is a group of units selected from a larger group (the population) to represent it, because the population is too large to study in its entirety. By studying the sample, the researcher might hope to draw valid conclusions about the larger group. The sample should, therefore, be representative of the general population. Since it is usually impractical to test every member of a population, a sample from the population is typically the best approach available. Before collecting the sample, it is important that the researcher accurately and completely defines the population, including a description of the members to be included (cf., "Stratified randomisation").

Sampling – Refers to the process of obtaining a sample.

Scatterplot – A scatterplot shows the scores on one variable plotted against scores on a second variable. Each value computed contributes one point to the scatterplot, on which points are plotted but not joined. The resulting pattern indicates the type and strength of the relationship between the two variables. A scatterplot is a useful summary of a set of data from two variables, gives a good visual picture of the relationship, and aids the interpretation of the correlation coefficient or regression model. Scatterplots should be presented when the relationship between two variables is of interest. Statistical summaries are no substitute for a full plot of the data. The plot below shows a very strong, but certainly not perfect, relationship between two variables (see also "Linearity"; "Correlation").



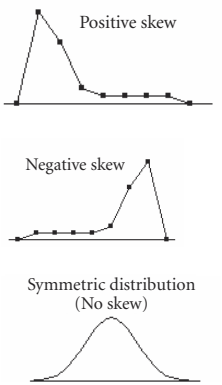
Scheffé's test – An example of a post-hoc test which is used to make unplanned comparisons among the means in an experiment. The Scheffé is one of the most powerful of such tests; for example, the combination of a one-way ANOVA and a Scheffé test will help the researcher identify whether there are significant differences in the means of different groups and pinpoint where those differences are really located.

Scores – Scores are the values given to subjects which then signify the position in which they lie along a scale associated with a specified characteristic.

SEE – The standard error of the estimate is a measure of the accuracy of predictions made

with a regression line.

Skew(ed) distribution – A distribution is said to be skewed when data plotted reveal that one of its tails is longer than the other. Positively-skewed data are graphically described in the first distribution below; there is a longer tail in the positive direction. In the middle graphic there is a longer tail described in the negative direction. Finally, the third distribution is symmetric and has no skew. It represents the familiar "bell-shaped curve" of the normal distribution (cf., "Normal distribution").



Slope – The slope of a regression line refers to its angle or steepness. Graphically, it is measured as the change in Y-axis values associated with a change of one unit on X-axis values. Lines with positive slopes are slanted up toward the right (small values on the X-axis align with small values on the Y-axis; large values on the X-axis align with large values on the Y-axis), while negative value slopes are slanted up toward the left (see also "Regression line").

Spearman correlation coefficient (ρ) – Commonly used procedures for making inferences about the population correlation coefficient, based on the "Pearson correlation coefficient", make the implicit assumption that the two variables are jointly normally distributed. When this assumption is not justified, a non-parametric measure such as the Spearman rank correlation coefficient might be more appropriate. Spearman's rho (ρ) may also be a better indicator that a relationship exists between two variables when that relationship is non-linear (cf., "Pearson correlation coefficient").

Split-half reliability – Split-half reliability statistics are simple measures of the internal consistency of a test obtained by dividing the test into two equal parts and calculating the correlation between scores on one half of the test with scores on the other half of the test (see also "Kuder-Richardson formulae").

Standard deviation (s or s.d.) – Standard deviation is the most-commonly used measure of the spread or dispersion of a set of data in inferential statistical procedures. It is calculated by taking the square root of the variance and is symbolised by "s.d", or "s". The more widely the values are spread out, the larger the standard deviation. In a normal distribution, about 68% of the scores are within one standard deviation (either side) of the mean and about 95% of the scores are within two standard deviations (cf., "Range"; "Dispersion"; "Variance").

Statistic(s) – The word "statistics" is used in several different senses. In the broadest sense, "statistics" refers to a range of techniques and procedures for analyzing, interpreting, and displaying data, and making decisions based on data. In a second usage, a "statistic" is defined as a numerical quantity (such as the mean) calculated in a sample. Such statistics are used to estimate parameters.

Statistical significance (level) – In hypothesis testing, the significance level is the criterion used for rejecting the null hypothesis. First, the difference between the results of the experiment and the null hypothesis is determined. Then, proceeding with the assumption that the null hypothesis is true, the probability (p) of a difference that large or larger is computed. Finally, this probability is compared to the significance level. If this probability is sufficiently

low (i.e., less than or equal to the significance level), then the null hypothesis is rejected and the outcome is said to be statistically significant. The researcher would want to make the significance level as small as possible in order to "protect" the null hypothesis and to avoid — as far as possible — inadvertently making false claims. The Greek letter alpha (α) is often used to indicate the significance level chosen (see also "Probability value/level"). An "alpha" of .01 (compared with .05 or .10) means the researcher is being relatively careful. He or she is only willing to risk being wrong 1 in a 100 times in rejecting the null hypothesis when it is true (i.e., saying there is an effect or relationship when there really is not).

Stratified randomisation – There may often be factors that divide up the population into sub-populations (groups / strata), and we may expect the data we are interested in to vary among these different sub-populations. This has to be accounted for when we select a sample from the population so that we obtain one that is representative of that population. Stratified sampling helps us achieve such an aim by taking samples from each stratum or sub-group of a population. The first step in such random sampling is to identify the stratification parameters of interest (e.g., only female L2 learning students between the ages of 18–21). Each stratification parameter represents a control variable. The study would then restrict the population to L2 learning students between these ages — as that is the specified control variable — and then sample across only female students — as this is the independent variable used. Any other strata would be similarly treated. Together with random selection within each stratum, stratification increases the chance that the sample will be representative of the population to whom we want to generalise any outcomes (see also "Sample").

Strength of association/relationship – (see "Effect size"; "Omega^2"; "Eta^2")

t test – A t-test is any of a number of tests based on the t distribution. The most common t-test is a test for a difference between two means.

Two-tailed hypothesis – (see "Directional hypothesis")

Type 1 and Type 2 errors – There are two kinds of errors that can be made in statistical significance testing: (1) a null hypothesis which is actually true can be incorrectly rejected, and (2) a false null hypothesis can fail to be rejected. The former error is called a Type 1 error and the latter a Type 2 error. A Type 2 error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost. More serious, however, is the case of a Type 1 error, since here a conclusion is drawn that the null hypothesis is false when, in fact, it is true. If a researcher sets up a test requiring strong evidence (i.e., setting the alpha level at .01 or .001) to reject the null hypothesis, it makes it unlikely that a true null hypothesis will be rejected (see also "Null hypothesis").

Variable – a property or quality of a person, piece of text, or object which is able, or seen, to differ or vary across these people, texts, or objects (see also "Independent variable"; "Dependent variable"; "Control variable"; "Moderator variable"; "Intervening variable").

Variability (see "Dispersion")

Variance – The variance is a measure of how spread out about its average value a distribution is. The larger the variance, the more scattered are the observations on average. It is calculated from the average squared deviation of each number from its mean (see also "Standard deviation"; "Range").

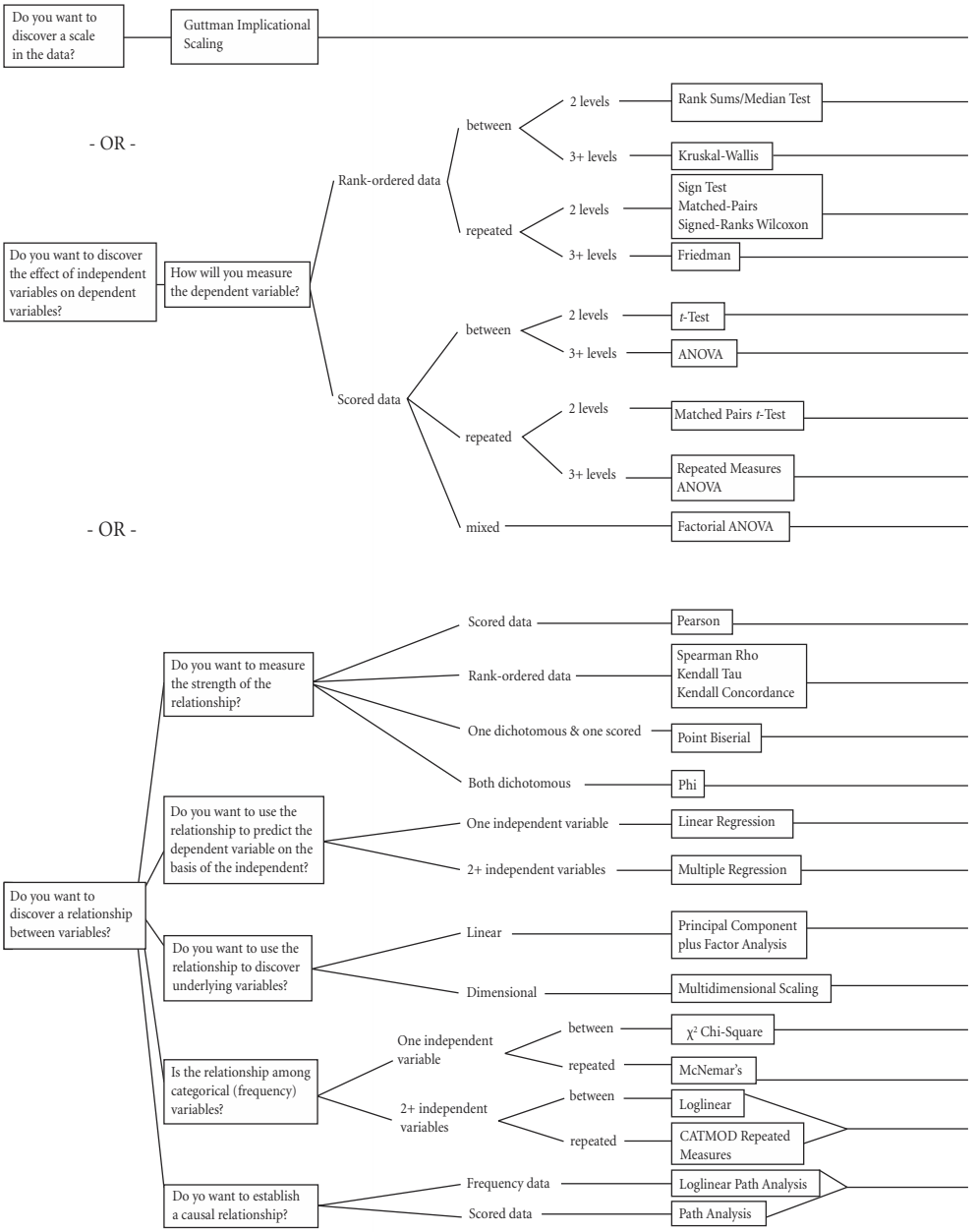
Within-group designs – Within-group/subject designs are those in which one or more of the independent variables are within-subject variables. These are often called repeated-measures designs since within-subjects variables always

involve taking repeated measurements from each subject. It is normal that subjects differ greatly in life. In between-groups designs, these differences among subjects are uncontrolled and are treated as error. In within-group designs, the same subjects are tested in each condition. Therefore, differences among subjects can be measured and separated from error. The distinction between within-group and between-groups designs will determine in part the choice of an appropriate statistical procedure for analysing the data (cf., "Between-groups designs"; "Mixed designs").

Appendices

Appendix I — Flow chart

Reprinted from from Hatch, E., and Lazaraton, A. 1991, The Research Manual. New York: Newbury House Publishers, pp. 544–545.


Do you want to discover a scale in the data?
Guttman Implicational Scaling
- OR -
Do you want to discover the effect of independent variables on dependent variables?
How will you measure the dependent variable?
Rank-ordered data
between
2 levels
Rank Sums/Median Test
3+ levels
Kruskal-Wallis
repeated
2 levels
Sign Test
Matched-Pairs
Signed-Ranks Wilcoxon
3+ levels
Friedman
Scored data
between
2 levels
t-Test
3+ levels
ANOVA
repeated
2 levels
Matched Pairs t-Test
3+ levels
Repeated Measures ANOVA
mixed
Factorial ANOVA
- OR -
Do you want to discover a relationship between variables?
Do you want to measure the strength of the relationship?
Scored data
Pearson
Rank-ordered data
Spearman Rho
Kendall Tau
Kendall Concordance
One dichotomous & one scored
Point Biserial
Both dichotomous
Phi
Do you want to use the relationship to predict the dependent variable on the basis of the independent?
One independent variable
Linear Regression
2+ independent variables
Multiple Regression
Do you want to use the relationship to discover underlying variables?
Linear
Principal Component plus Factor Analysis
Dimensional
Multidimensional Scaling
Is the relationship among categorical (frequency) variables?
One independent variable
between
χ² Chi-Square
repeated
McNemar's
2+ independent variables
between
Loglinear
repeated
CATMOD Repeated Measures
Do yo want to establish a causal relationship?
Frequency data
Loglinear Path Analysis
Scored data
Path Analysis

How will you interpret the results?

Check that the coefficient of scalability is over .60	Interpret scale in light of reasonableness of the cutoff point, number of instances required, missing values, and context used to elicit forms.
Compare the groups Check *z* value	If the *z* score is significant, then one group has more *Ss* in higher ranks than the other. Use eta^2 for strength of association.
Compare the groups Check *H* value	If *H* is significant, the groups differ; use Ryan's procedure to locate which groups differ. Use eta^2 for strength of association.
Compare the groups Check *R* or *z* value	If the Sign test *R* or Wilcoxon *z* is significant, there is a change from time 1 to time 2. Use eta^2 for strength of association for Wilcoxon.
Compare the groups Check χ^2	If χ^2 is significant, there is a change over several time points (or msrs.) To locate the difference more precisely report the results of the Nemenyi's procedure. Use eta^2 for strength of association.
Compare the two means Check *t* value	If the *t* value is significant, the two groups differ. Use eta^2 for strength of association.
Compare 3+ means Check *F* ratio	If the *F* ratio is significant, the groups differ. To locate the differences more precisely, interpret the multiple-range test (Scheffé, Tukey, or Newman-Keuls). Use $omega^2$ or eta^2 for strength of association.
Compare the two means Check *t* value	If the *t* value is significant, there is a difference in the means for the two times or measures. Use eta^2 for strength of association.
Compare 3+ means Check *F* ratio	If the *F* ratio is significant, the same (or matched) *Ss* perform differently on repeated measures. Use a multiple-range test to locate precise differences. Use eta^2 for strength of association.
Compare the means Check *F* ratios	Step 1. If the interaction is significant, chart the means to show the interaction and interpret it. Interpret main effects in light of the interaction. Step 2. If the interaction is not significant, interpret the difference in the main effects. Use a multiple-range test to locate precise differences. Use eta^2 to show the strength of association.
Check the strength of each correlation	r^2 shows the amount of overlap between each pair of variables. Be sure to correct for attenuation if measures are not of equel reliability.
Check the probability of the correlation	If the correlation is significant, it shows that the *H* of no relation can be rejected. Interpret the value "sensibly" in terms of strength of relationship.
Check the value of the correlation	Explain the correlation in a "sensible" way.
Check χ^2 for significance	Explain the correlation in a "sensible" way.
Report predicted scores Check the SEE	The stronger the correlation and the smaller the SEE, the better the prediction will be.
Check each added variable	Identify the first independent variable, then the overlap of the second with the first to see how much each contributes (as well as their joint contribution) to explain variance in the dependent variable. Explain how much additional information is given by each succeeding independent variable.
Check each factor loading	If possible, once the number of factors has been determined, label each factor by consulting variables with high loadings vs. variables with low loadings on each. Else, label them as factor A, B, C, etc.
Check solutions and stress	Once a solution (about number of dimensions) has been identified or selected, label each dimension by consulting items in the cluster and those distant from cluster. Else, label them as dimension A, B, C, etc.
Check χ^2 value & $(O-E)^2/E$ values	If χ^2 is significant, the distribution differs from the expected distribution. Show which cells differ most from expected cell frequency or do a Ryan's procedure to locate the difference more precisely. Use Phi or Cramers V for strength of association.
Check z value	If *z* is significant, conclude there is a change in proportion of *Ss* from time 1 to time 2.
Check parameter estimates to reduce model, compare models	The parameter estimates show which interactions and main effects are significant. To pare the model, compare various models with the saturated model. Decisions should be based on statistical and substantive arguments.
Check the paths to see which can be trimmed from the model	Use the analysis to trim paths from the model. Interpret the findings on both statistical and substantive grounds.

Appendix II — Table of assumptions for popular statistical tests

Adapted from Brown, J. D. 1992. Statistics as a foreign language: Part 2. *Tesol Quarterly*, 26, 4, pp. 629–664.

Statistical procedure/ Assumptions	Independence of groups	Independence of observations	Normality	Equal variances	Linearity	Non-multicolinearity	Homoscedasticity	Other assumptions
Correlation								
Pearson r		●	●		●		●	
Spearman rho		●						
Kendall tau		●						
Kendall W		●						
Point-biserial correlation		●			●			
Phi coefficient		●	●		●			
Correlation/prediction								
Simple regression		●	●		●		●	
Multiple regression		●	●		●	●	●	
Loglinear analysis		●						No more than 20% of expected frequencies less than or equal to 5
Group differences								
z statistic (large samples)	●	●	●	●				
t test (any samples)	●	●	●	●				
One-way ANOVA	●	●	●	●				
One-way ANCOVA	●	●	●	●	●	●		
Matched pairs t-test		●	●	●				
Repeated measures ANOVA			●	●				

Statistical procedure/ Assumptions	Independence of groups	Independence of observations	Normality	Equal variances	Linearity	Non-multicolinearity	Homoscedasticity	Other assumptions
Repeated measures ANCOVA			●	●	●	●		
n-way ANOVA	●	●	●	●				
n-way ANCOVA	●	●	●	●	●	●		
n-way repeated measures ANOVA			●	●				
n-way repeated measures ANCOVA			●	●	●	●		
Multivariate ANOVA	●	●	●	●	●			
Multivariate ANCOVA	●	●	●	●	●	●		
Multivariate n-way ANOVA	●	●	●	●	●			
Multivariate n-way ANCOVA	●	●	●	●	●	●		
Median test	●	●						
Mann U/Wilcoxon	●	●						
Kruskal-Wallis	●	●						
Sign test	●	●						
Friedman One-way ANOVA	●	●						
Frequencies								
Chi-squared	●	●						Expected frequencies greater or equal to 5 if the df is greater or equal to 2; greater than or equal to 10 if the df equals 1.
McNemar test								Differences all in same direction (same sign)
Fisher's exact test	●	●						
n-way chi-squared	●	●						
Exploratory statistics								

Statistical procedure/ Assumptions	Independence of groups	Independence of observations	Normality	Equal variances	Linearity	Non-multicolinearity	Homoscedasticity	Other assumptions
Principal component analysis			●		●		●	Factorability of R
Factor analysis			●		●	●	●	Factorability of R
Multidimensional scaling			●		●	●		
Cluster analysis			●		●	●		
One-way discriminant analysis			●		●	●		Homogeneity of variance-covariance matrices
n-way discriminant analysis			●		●	●		Homogeneity of variance-covariance matrices
Guttman scaling								Scalable and reproducible
Path analysis			●		●	●	●	All relevant variables included; variables are causal

Appendix III — Useful statistical tables

The distribution of the *F*-statistic (.05)

	***df* for greater mean square**											
df	**1**	**2**	**3**	**4**	**5**	**6**	**8**	**10**	**15**	**25**	**50**	**100**
1	161.5	199.5	215.8	224.6	230.2	234.0	238.9	241.9	246.0	249.3	251.8	253.1
2	18.51	19.00	19.16	19.25	19.30	19.31	19.37	19.40	19.43	19.456	19.476	19.49
3	10.1280	9.5521	9.2766	9.1172	9.0134	8.9407	8.8452	8.7855	8.7028	8.6341	8.5810	8.5539
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0410	5.9644	5.8578	5.7687	5.6995	5.6640
5	6.6079	5.7861	5.4094	5.1922	5.0503	4.9503	4.8183	4.7351	4.6188	4.5209	4.4444	4.4051
6	5.9874	5.1432	4.7571	4.5337	4.3874	4.2839	4.1468	4.0600	3.9381	3.8348	3.7537	3.7117

	df for greater mean square											
df	1	2	3	4	5	6	8	10	15	25	50	100
7	5.5915	4.7374	4.3468	4.1203	3.9715	3.8660	3.7257	3.6365	3.5107	3.4036	3.3189	3.2749
8	5.3176	4.4590	4.0662	3.8379	3.6875	3.5806	3.4381	3.3472	3.2184	3.1081	3.0204	2.9747
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2296	3.1373	3.0061	2.8932	2.8028	2.7556
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.0717	2.9782	2.8450	2.7298	2.6371	2.5884
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	2.9480	2.8536	2.7186	2.6014	2.5066	2.4566
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.8486	2.7534	2.6169	2.4977	2.4010	2.3498
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.7669	2.6710	2.5331	2.4123	2.3138	2.2614
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.6987	2.6022	2.4630	2.3407	2.2405	2.1870
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.6408	2.5437	2.4034	2.2797	2.1780	2.1234
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.5911	2.4935	2.3522	2.2272	2.1240	2.0685
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.5480	2.4499	2.3077	2.1815	2.0769	2.0204
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5102	2.4117	2.2686	2.1413	2.0354	1.9780
19	4.3808	3.5219	3.1274	2.8951	2.7401	2.6283	2.4768	2.3779	2.2341	2.1057	1.9986	1.9403
20	4.3513	3.4928	3.0984	2.8661	2.7109	2.5990	2.4471	2.3479	2.2033	2.0739	1.9656	1.9066
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4205	2.3210	2.1757	2.0454	1.9360	1.8761
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.3965	2.2967	2.1508	2.0196	1.9092	1.8486
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.3748	2.2747	2.1282	1.9963	1.8848	1.8234
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.3551	2.2547	2.1077	1.9750	1.8625	1.8005
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.3371	2.2365	2.0889	1.9554	1.8421	1.7794
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3205	2.2197	2.0716	1.9375	1.8233	1.7599
27	4.2100	3.3541	2.9603	2.7278	2.5719	2.4591	2.3053	2.2043	2.0558	1.9210	1.8059	1.7419
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.2913	2.1900	2.0411	1.9057	1.7898	1.7251
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.2782	2.1768	2.0275	1.8915	1.7748	1.7096
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.2662	2.1646	2.0148	1.8782	1.7609	1.6950
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.1802	2.0773	1.9245	1.7835	1.6600	1.5892
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.0970	1.9926	1.8364	1.6902	1.5590	1.4814
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0164	1.9105	1.7505	1.5980	1.4565	1.3685
1000	3.8508	3.0047	2.6138	2.3808	2.2231	2.1076	1.9476	1.8402	1.6764	1.5171	1.3632	1.2596

Distribution of the *F*-statistic (.01)

	df for greater mean square											
df	1	2	3	4	5	6	8	10	15	25	50	100
1	4052.2	4999.3	5403.5	5624.3	5764.0	5859.0	5981.0	6055.9	6157.0	6239.9	6302.3	6333.9
2	98.5019	99.0003	99.1640	99.2513	99.3023	99.3314	99.3750	99.3969	99.4332	99.4587	99.4769	99.4914
3	34.1161	30.8164	29.4567	28.7100	28.2371	27.9106	27.4895	27.2285	26.8719	26.5791	26.3544	26.2407

	df for greater mean square											
df	1	2	3	4	5	6	8	10	15	25	50	100
4	21.1976	17.9998	16.6942	15.9771	15.5219	15.2068	14.7988	14.5460	14.1981	13.9107	13.6897	13.5769
5	16.2581	13.2741	12.0599	11.3919	10.9671	10.6722	10.2893	10.0511	9.7223	9.4492	9.2377	9.1300
6	13.7452	10.9249	9.7796	9.1484	8.7459	8.4660	8.1017	7.8742	7.5590	7.2960	7.0914	6.9867
7	12.2463	9.5465	8.4513	7.8467	7.4604	7.1914	6.8401	6.6201	6.3144	6.0579	5.8577	5.7546
8	11.2586	8.6491	7.5910	7.0061	6.6318	6.3707	6.0288	5.8143	5.5152	5.2631	5.0654	4.9633
9	10.5615	8.0215	6.9920	6.4221	6.0569	5.8018	5.4671	5.2565	4.9621	4.7130	4.5167	4.4150
10	10.0442	7.5595	6.5523	5.9944	5.6364	5.3858	5.0567	4.8491	4.5582	4.3111	4.1155	4.0137
11	9.6461	7.2057	6.2167	5.6683	5.3160	5.0692	4.7445	4.5393	4.2509	4.0051	3.8097	3.7077
12	9.3303	6.9266	5.9525	5.4119	5.0644	4.8205	4.4994	4.2961	4.0096	3.7647	3.5692	3.4668
13	9.0738	6.7009	5.7394	5.2053	4.8616	4.6203	4.3021	4.1003	3.8154	3.5710	3.3752	3.2723
14	8.8617	6.5149	5.5639	5.0354	4.6950	4.4558	4.1400	3.9394	3.6557	3.4116	3.2153	3.1118
15	8.6832	6.3588	5.4170	4.8932	4.5556	4.3183	4.0044	3.8049	3.5222	3.2782	3.0814	2.9772
16	8.5309	6.2263	5.2922	4.7726	4.4374	4.2016	3.8896	3.6909	3.4090	3.1650	2.9675	2.8627
17	8.3998	6.1121	5.1850	4.6689	4.3360	4.1015	3.7909	3.5931	3.3117	3.0676	2.8694	2.7639
18	8.2855	6.0129	5.0919	4.5790	4.2479	4.0146	3.7054	3.5081	3.2273	2.9831	2.7841	2.6779
19	8.1850	5.9259	5.0103	4.5002	4.1708	3.9386	3.6305	3.4338	3.1533	2.9089	2.7092	2.6023
20	8.0960	5.8490	4.9382	4.4307	4.1027	3.8714	3.5644	3.3682	3.0880	2.8434	2.6430	2.5353
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.5056	3.3098	3.0300	2.7850	2.5838	2.4755
22	7.9453	5.7190	4.8166	4.3134	3.9880	3.7583	3.4530	3.2576	2.9779	2.7328	2.5308	2.4218
23	7.8811	5.6637	4.7648	4.2635	3.9392	3.7102	3.4057	3.2106	2.9311	2.6857	2.4829	2.3732
24	7.8229	5.6136	4.7181	4.2185	3.8951	3.6667	3.3629	3.1681	2.8887	2.6430	2.4395	2.3291
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.3239	3.1294	2.8502	2.6041	2.3999	2.2888
26	7.7213	5.5263	4.6365	4.1400	3.8183	3.5911	3.2884	3.0941	2.8150	2.5686	2.3637	2.2519
27	7.6767	5.4881	4.6009	4.1056	3.7847	3.5580	3.2558	3.0618	2.7827	2.5360	2.3304	2.2180
28	7.6357	5.4529	4.5681	4.0740	3.7539	3.5276	3.2259	3.0320	2.7530	2.5060	2.2997	2.1867
29	7.5977	5.4205	4.5378	4.0449	3.7254	3.4995	3.1982	3.0045	2.7256	2.4783	2.2713	2.1577
30	7.5624	5.3903	4.5097	4.0179	3.6990	3.4735	3.1726	2.9791	2.7002	2.4526	2.2450	2.1307
40	7.3142	5.1785	4.3126	3.8283	3.5138	3.2910	2.9930	2.8005	2.5216	2.2714	2.0581	1.9383
60	7.0771	4.9774	4.1259	3.6491	3.3389	3.1187	2.8233	2.6318	2.3523	2.0984	1.8772	1.7493
120	6.8509	4.7865	3.9491	3.4795	3.1735	2.9559	2.6629	2.4721	2.1915	1.9325	1.7000	1.5592
1000	6.6603	4.6264	3.8012	3.3380	3.0356	2.8200	2.5290	2.3386	2.0565	1.7915	1.5445	1.3835

Distribution of the *t*-statistic

	Level of significance for one-tailed tests					
	0.100	0.050	0.025	0.010	0.005	0.0005
	Level of significance for two-tailed tests					
df	0.200	0.100	0.050	0.020	0.010	0.001

1	3.0777	6.3138	12.7061	31.8202	63.6568	636.6409
2	1.8856	2.9200	4.3026	6.9646	9.9248	31.5971
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9250
4	1.5332	2.1318	2.7764	3.7470	4.6041	8.6097
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8686
6	1.4398	1.9432	2.4469	3.1427	3.7075	5.9587
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5868
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370
12	1.3562	1.7823	2.1788	2.6810	3.0546	4.3179
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208
14	1.3450	1.7613	2.1448	2.6245	2.9769	4.1403
15	1.3406	1.7530	2.1314	2.6025	2.9467	4.0728
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9650
18	1.3304	1.7341	2.1009	2.5524	2.8785	3.9217
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8835
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8496
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8192
22	1.3212	1.7171	2.0739	2.5083	2.8187	3.7922
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7677
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7252
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6738
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460
40	1.3031	1.6838	2.0211	2.4233	2.7045	3.5509
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.4601
120	1.2886	1.6577	1.9799	2.3578	2.6174	3.3734
1000	1.2824	1.6464	1.9623	2.3301	2.5808	3.3003

To determine if your calculated value of t is statistically significant: (1) Determine if you are working with a one-tailed or two-tailed t-test, (2) find the appropriate probability level column, (3) find appropriate df (generally n-1), and then (4) find the critical value in the body of the table. Now, (5) compare your calculated value with the table value above. Your calculated value must be equal to or greater than the table value to be considered statistically significant at the significance level noted above.

Critical values of *U* for the Mann-Whitney test

p-value = 0.01 for one-tailed test, 0.02 for two-tailed test

N1 \ N2	9	10	11	12	13	14	15	16	17	18	19	20
3	1	1	1	2	2	2	3	3	4	4	4	5
4	3	3	4	5	5	6	7	7	8	9	9	10
5	5	6	7	8	9	10	11	12	13	14	15	16
6	7	8	9	11	12	13	15	16	18	19	20	22
7	9	11	12	14	16	17	19	21	23	24	26	28
8	11	13	15	17	20	22	24	26	28	30	32	34
9	14	16	18	21	23	26	28	31	33	36	38	40
10	16	19	22	24	27	30	33	36	38	41	44	47
11	18	22	25	28	31	34	37	41	44	47	50	53
12	21	24	28	31	35	38	42	46	49	53	56	60
13	23	27	31	35	39	43	47	51	55	59	63	67
14	26	30	34	38	43	47	51	56	60	65	69	73
15	28	33	37	42	47	51	56	61	66	70	75	80

p-value = 0.05 for the one-tailed test, 0.1 for the two-tailed test

N1 \ N2	9	10	11	12	13	14	15	16	17	18	19	20
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100

For any N1 and N2 the observed value of U is significant if it is equal to or less than the critical values shown.

Distribution of chi squared

	Significance													
df	**0.99**	**0.98**	**0.95**	**0.9**	**0.8**	**0.7**	**0.5**	**0.3**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.001**
1	.000	.001	.004	.016	.064	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.828
2	.020	.040	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.816
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.458
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.124
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.892
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.301
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703

To determine if your calculated value of chi-squared is statistically significant: (1) find the appropriate probability level column, (2) find appropriate df, and then (3) find the critical value in the body of the table. Now, (4) compare your calculated value with the table value above. Your calculated value must be equal to or greater than the table value to be considered statistically significant at the significance level noted above.

Critical values of the Pearson product-moment correlation

If the observed value of r is greater than or equal to the tabled value for the desired level of significance and degrees of freedom (number of pairs of scores minus 2), we conclude that a statistically significant relationship between the variables does exist in the population sampled.

	Level of significance for a nondirectional (two-tailed) test				
df $N-2$	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
1	0.9877	0.9969	0.9995	0.9999	1.0000
2	0.9000	0.9500	0.9800	0.9900	0.9990
3	0.8054	0.8783	0.9343	0.9587	0.9912
4	0.7293	0.8114	0.8822	0.9172	0.9741
5	0.6694	0.7545	0.8329	0.8745	0.9507
6	0.6215	0.7067	0.7887	0.8343	0.9249
7	0.5822	0.6664	0.7498	0.7977	0.8982
8	0.5494	0.6319	0.7155	0.7646	0.8721
9	0.5214	0.6021	0.6851	0.7348	0.8471
10	0.4973	0.5760	0.6581	0.7079	0.8233
11	0.4762	0.5529	0.6339	0.6835	0.8010
12	0.4575	0.5324	0.6120	0.6614	0.7800
13	0.4409	0.5139	0.5923	0.6411	0.7603
14	0.4259	0.4973	0.5742	0.6226	0.7420
15	0.4124	0.4821	0.5577	0.6055	0.7246
16	0.4000	0.4683	0.5425	0.5897	0.7084
17	0.3887	0.4555	0.5285	0.5751	0.6932
18	0.3783	0.4438	0.5155	0.5614	0.6787
19	0.3687	0.4329	0.5034	0.5487	0.6652
20	0.3598	0.4227	0.4921	0.5368	0.6524
25	0.3233	0.3809	0.4451	0.4869	0.5974
30	0.2960	0.3494	0.4093	0.4487	0.5541
35	0.2746	0.3246	0.3810	0.4182	0.5189
40	0.2573	0.3044	0.3578	0.3932	0.4896
45	0.2428	0.2875	0.3384	0.3721	0.4648
50	0.2306	0.2732	0.3218	0.3541	0.4433
60	0.2108	0.2500	0.2948	0.3248	0.4078
70	0.1954	0.2319	0.2737	0.3017	0.3799
80	0.1829	0.2172	0.2565	0.2830	0.3568
90	0.1726	0.2050	0.2422	0.2673	0.3375

	Level of significance for a nondirectional (two-tailed) test				
df $N-2$	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
100	0.1638	0.1946	0.2301	0.2540	0.3211

Critical values of the Spearman rank-order correlation

If the observed value of r is greater than or equal to the tabled value for the desired level of significance and number of pairs, we conclude that a statistically significant relationship between these variables does exist in the population sampled.

	Level of significance for two-tailed test			
N*	**0.10**	**0.05**	**0.02**	**0.01**
5	0.900	1.000	1.000	–
6	0.829	0.886	0.943	1.000
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.600	0.683	0.783	0.833
10	0.564	0.648	0.746	0.794
12	0.506	0.591	0.712	0.777
14	0.456	0.544	0.645	0.715
16	0.425	0.506	0.601	0.665
18	0.399	0.475	0.564	0.625
20	0.377	0.450	0.534	0.591
22	0.359	0.428	0.508	0.562
24	0.343	0.409	0.485	0.537
26	0.329	0.392	0.465	0.515
28	0.317	0.377	0.448	0.496
30	0.306	0.364	0.432	0.478

* N = number of pairs

Further reading

Allwright, R. *Observation in the Language Classroom*. New York: Longman, 1988.

Aron, A., and E. Aron. *Statistics for psychology*. Upper Saddle River, NJ: Prentice Hall, 1999.

Aronson, E., P. Ellsworth, J. Carlsmith, and M. Gonzales. *Methods of research in social psychology*. New York: McGraw Hill, 1990.

Blaxter, L., C. Hughes, and M. Tight. *How to Research*. Buckingham: Open University Press, 1996.

Borg, W. *Educational Research: an Introduction*. New York: Addison Wesley Longman, 1989.

Borg, W., and M. Gall. *Educational Research: an Introduction*. New York: Longman, 1979.

Brown, J.D. *Understanding Research in Second Language Learning*. Cambridge: Cambridge University Press, 1988.

Campbell, D., and J. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1966.

Cochran, W. *Statistical Analysis in Psychology and Education*. New York: McGraw Hill, 1981.

Cresswell, J. *Research Design: Qualitative and Quantitative Approaches*. Thousand Oaks: Sage Publications, 1994.

Dickinson Gibbons. J. *Non-Parametric Statistics: an Introduction*. Thousand Oaks: Sage Publications, 1994.

Greene, J., and M. D'Oliveira. *Learning to Use Statistical Tests in Psychology*. Milton Keynes: Open University Press, 1982.

Guilford J., and B. Fruchter. *Fundamental Statistics in Psychology and Education*. New York: McGraw Hill, 1973.

Hart, C. *Doing a Literature Review*. London: SAGE Publications, 1998.

Hatch, E., and H. Farhady. *Research Design and Statistics for Applied Linguistics*. Rowley, Mass.: Newbury House, 1982.

Hatch, E., and A. Lazaraton. *Design and Statistics for Applied Linguistics*. New York: Newbury House Publishers, 1991.

Hays, W. *Statistics for the Social Sciences*. New York: Holt, Rinehardt and Winston, 1973.

Hollander, M., and A. Wolfe. *Non-parametric Statistical Methods*. New York: J. Wiley, 1973.

Hopkins, D. *A Teacher's Guide to Classroom Research*. Buckingham: Open University Press, 1993.

Johnson, D. *Approaches to Research in Second Language Learning*. New York: Addison Wesley Longman, 1992.

Kerlinger, F. *Foundations of Behavioural Research*. New York: Holt, Rinehardt and Winston, 1973.

Leong, F., and J. Austin. (eds.) *The Psychology Research Handbook.* Thousand Oaks: Sage Publications, 1996.

Lewis-Beck, M. (ed.) *Experimental Design and Methods.* Thousand Oaks: Sage Publications, 1993.

Mehrens, W., and I. Lehmann. *Measurement and Evaluation in Education and Psychology.* New York: Holt, Rinehart, and Winston, 1969.

Sapsford, R., and V. Jupp. (eds.) *Data Collection and Analysis.* Thousand Oaks: Sage Publications, 1996.

Scholfield, P. *Quantifying Language.* Clevedon: Multilingual Matters, 1995.

Seliger, H., and E. Shohamy. *Second Language Research Methods.* Oxford: Oxford University Press, 1990.

Shaughnessy, J. J., E. Zechmeister, and J. S. Shaughnessy. *Research Methods in Psychology.* New York: McGraw Hill, 2000.

Sowell, E. *Educational Research: an Integrative Introduction.* New York: McGraw Hill, 2000.

Spector, P. *Research Designs.* Thousand Oaks: Sage Publications, 1994.

Sternberg, R. *The psychologist's companion: a guide to scientific writing for students and researchers.* New York: Cambridge University Press, 1993.

Tarone, E., S. Gass, and A. Cohen. (eds.) *Research Methodology in Second Language Acquisition.* Hillsdale, NJ: Erlbaum, 1994.

Thorndike, R. *Correlational Procedures for Research.* New York: Gardner Press, 1978.

Tuckman, B. *Conducting Educational Research.* New York: Harcourt Brace College, 1994.

Vockell, E. *Educational Research.* New York: Macmillan, 1979.

Wiersma, W. *Research Methods in Education.* Boston: Allyn and Bacon, 1985.

Woods, A., P. Fletcher, and A. Hughes. *Statistics in Language Studies.* Cambridge: Cambridge University Press, 1993.

Wright, D. *Understanding Statistics.* Thousand Oaks: Sage Publications, 1996.

Index of main subjects in Textbook

In the series LANGUAGE LEARNING & LANGUAGE TEACHING (LL&LT) the following titles have been published thus far, or are scheduled for publication:

1. CHUN, Dorothy M.: *Discourse Intonation in L2. From theory and research to practice.* 2002.
2. ROBINSON, Peter (ed.): *Individual Differences and Instructed Language Learning*. 2002.
3. PORTE, Graeme Keith: *Appraising Research in Second Language Learning. A practical approach to critical analysis of quantitative research.* 2002.
4. TRAPPES-LOMAX, Hugh and Gibson FERGUSON: *Language in Language Teacher Education.* n.y.p.
5. GASS, Susan, Kathleen BARDOVI-HARLIG, Sally Sieloff MAGNAN and Joel WALZ (eds.): *Pedagogical Norms for Second and Foreign Language Learning and Teaching.* 2002.
6. GRANGER, Sylviane, Joseph HUNG and Stephanie PETCH-TYSON (eds.): *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* n.y.p.