A framework for the use and interpretation of statistics in reading instruction

Jeanette Brits

M.A.

Thesis submitted for the degree Philosophiae Doctor in English at the North-West University (Potchefstroom Campus)

Promoter: Prof. C. Nel

Assistant Promoter: Prof. H.S. Steyn

2007

Potchefstroom

ACKNOWLEDGEMENTS

I would like to thank the following individuals and concerns without whose cooperation this research would not have materialised;

- Prof. C. Nel, my promoter, for her support, prompt feedback, and expert guidance.
- Prof. H.S. Steyn, my assistant promoter, for his valuable statistical guidance.
- My husband, children and parents, for the sacrifices they made, their love and encouragement.
- North-West University (Potchefstroom Campus) for financial assistance.
- The Director of the Focus Area 04 for financial assistance.
- Finally, all praise belongs to God, for His love and omniscient guidance.

TABLE OF CONTENTS

CHAPTER 3

STATISTICAL SIGNIFICANCE TESTING

CHAPTER 4

ALTERNATIVE ANALYSIS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

Keywords: reading instruction research; experimental studies; randomised control trials; intervention; programmes; national reading panel; statistical significance testing

There are few instructional tasks more important than teaching children to read. The consequences of low achievement in reading are costly both to individuals and society. Low achievement in literacy correlates with high rates of school drop-out, poverty, and underemployment. The far-reaching effects of literacy achievement have heightened the interest of educators and non-educators alike in the teaching of reading. Successful efforts to improve reading achievement emphasise identification and implementation of *evidence-based* practices that promote high rates of achievement when used in classrooms by teachers with diverse instructional styles with children who have diverse instructional needs and interests.

Being able to recognise what characterises rigorous evidence-based reading instruction is essential to choosing the right reading curriculum (i.e., what method or approach). It will be necessary to ensure that general classroom reading instruction is of universally high quality and that practitioners are prepared to effectively implement validated reading interventions. When educators are not familiar with research methodologies and findings, national and provincial departments of education may find themselves implementing fads or incomplete findings.

The choice of method of instruction is very often based on empirical research studies. The selection of statistical procedures is an integral part of the research process. Statistical significance testing is a prominent feature of data analysis in language learning studies and also specifically, reading instruction studies.

For many years, methodologists have debated what statistical significance testing means and how it should be used in the interpretation of substantive results. Researchers have long placed a premium on the use of statistical significance testing. However, criticisms of the statistical significance testing procedure are prevalent and occur across many scientific disciplines.

Critics of statistical significance tests have made several suggestions, with the underlying theme being for researchers to examine and interpret their data carefully and thoroughly, rather than relying solely upon p values in determining which results are important enough to examine further and report in journals. Specific suggestions include the use of effect sizes, confidence intervals, and power.

The purpose of this study was to:

- determine what the state of affairs is with regard to statistical significance testing in reading instruction research, with specific reference to post-1999 literature (post-1999 literature was selected because of the specific request, made by Wilkinson and the Task Force on Statistical Inference in 1999, to include the reporting of effect sizes in empirical research studies);
- determine what the criticisms as well as the defences are that have been offered for statistical significance testing;
- determine what the alternatives or supplements are to statistical significance testing in reading instruction research;
- To provide a framework for the most effective and appropriate selection, use and representation of statistical significance testing in the reading instruction research field.

A comprehensive survey on the use of statistical significance testing, as manifested in randomly selected journals, was undertaken. Six journals (i.e., *System, Language Learning and Technology, The Reading Matrix, Scientific Studies of Reading, Teaching English as a Second or Foreign Language (TESL-EJ); South African Journal for Language Teaching*) regularly including articles related to reading instruction research and publishing articles reporting

statistical analyses, were reviewed and analysed. All articles in these journals from 2000-2005, employing statistical analyses were reviewed and analysed. The data was analysed by means of descriptive statistics (i.e., frequency counts and percentages). Qualitative reporting was also utilized.

A review of six readily accessible (online) journals publishing research on reading instruction indicated that researchers/authors rely very heavily on statistical significance testing and very seldom, if ever, report effect size/effect magnitude or confidence interval measures when documenting their results.

A review of the literature indicates that null hypothesis significance testing has been and is a controversial method of extracting information from experimental data and of guiding the formation of scientific conclusions. Several alternatives or complements to null hypothesis significance testing, namely effect sizes, confidence intervals and power analysis have been suggested.

The following central theoretical statement was formulated for this study:

> Statistical significance tests should be supplemented with accurate reports of effect size, power analyses and confidence intervals in reading research studies. In addition, quantitative studies, utilising statistics as stated in the previous sentence, should be supplemented with qualitative studies in order to obtain a more comprehensive picture of reading instruction research.

Research indicates that no single study ever establishes a programme or practice as effective; moreover it is the convergence of evidence from a variety of study designs that is ultimately scientifically convincing. When evaluating studies and claims of evidence, educators must not determine whether the study is quantitative or qualitative in nature, but rather if the study meets the standards of scientific research.

The proposed framework presented in this study consists of three main parts, namely, part one focuses on the study's description of the intervention and the

random assignment process, part two focuses on the study's collection of data and part three focuses on the study's reporting of results, specifically the statistical reporting of the results.

OPSOMMING

Sleutelwoorde: leesonderrignavorsing: eksperimentele studies; ewekansige-kontrole proeflopies; ingryping; programme; nasionale leespaneel; statistiese-betekenisvolle toetsing.

Baie min onderrigtake is belangriker as om kinders te leer om te lees. Swak leesvermoë plaas 'n groot las op die individu sowel as die samelewing. Lae prestasie in geletterdheid korreleer met voortydige skoolverlating, armoede en werkloosheid. Die verreikende gevolge van geletterdheidsprestasie het die verbeelding van opvoeders sowel as nie-opvoeders ten opsigte van leesonderrig aangegryp. Suksesvolle pogings om leespretasie te verbeter, beklemtoon die identifikasie en implementering van getuienisgebaseerde praktyke wat bevorderlik is vir hoë prestasievlakke wanneer dit deur onderwysers met diverse onderrigstyle toegepas word op kinders met uiteenlopende onderrigbehoeftes en –belangstellings.

Die vermoë om die eienskappe van streng getuienisgebaseerde leesonderrig te identifiseer is noodsaaklik wanneer 'n leeskurrikulum gekies word (bv. watter metode of benadering). Die sal noodsaaklik wees om te verseker dat algemene klaskamer leesonderrig van universele hoë gehalte is en dat praktisyns bereid is om geldige leesintervensies te implementeer. As opvoeders nie op hoogte is met navorsingsmetodologie en bevindings nie, mag nasionale- en provinsiale onderwysdepartemente moontlik giere of onvolledige bevindings implementeer.

Die keuse van die onderrig metode word dikwels gebaseer op empiriese navorsingstudies. Die keuse van statistiese prosedures is 'n integrale deel van die navorsingsproses. Statistiese-betekenisvolle toetsing is 'n prominente eienskap van data-analise in taal-aanleerstudies en meer spesifiek, in leesonderrigstudies.

Metodologiste debatteer al vir baie jare oor wat statisties-betekenisvolle toetsing nou eintlik is en hoe dit behoort gebruik te word in die interpretasie van onafhanklike resultate. Navorsers het lankal voorkeur gegee aan die gebruik van statisties-betekenisvolle toetsing, alhoewel daar tog vanuit verskeie wetenskaplike dissiplines ook sterk stemme van kritiek daarteen opgaan.

Kritici van statistiese-betekenisvolle toetsing het heelwat voorstelle gemaak wat hoofsaaklik daarop neerkom dat navorsers hulle data noukeurig moet nagaan en interpreteer, eerder as om slegs op p-waardes staat te maak wanneer besluit word watter gegewens belangrik genoeg is om verder te bestudeer en om in vaktydskrifte te publiseer. Spesifieke voorstelle sluit in die gebruik van effekgroottes, vertroue-intervalle en mag.

Die doel van hierdie studie is om:

- te bepaal wat die toedrag van sake ten opsigte van statisties-betekenisvolle toetsing in leesonderrig is met spesifieke verwysing na post-1999 literatuur (post-1999 literatuur is gekies as gevolg van 'n spesifieke versoek deur Wilkinson en die Taakmag op Statistiese Inferensie van 1999, om rapportering oor effekgroottes in empiriese navorsingstudies in te sluit);
- te bepaal wat die punte van kritiek teen en die punte ten gunste van statisties-betekenisvolle toetsing is, wat al gepubliseer is;
- te bepaal watter alternatiewe vir of aanvullings tot statistiese-betekenisvolle toetsing in leesonderrignavorsing al gepubliseer is;
- 'n raamwerk daar te stel vir die mees effektiewe en gepaste keuse-, aanwending- en voorstelling van statistiese-betekenisvolle toetsing in die navorsingsveld van leesonderrig.

'n Omvattende opname oor die gebruik van statisties-betekenisvolle toetsing soos wat die voorkom in ewekansig-gekose publikasies, is uitgevoer. Ses publikasies (d.i., *System, Language Learning and Technology, The Reading Matrix, Scientific Studies of Reading, Teaching English as a Second or*

Foreign Language (TESL-EJ); South African Journal for Language Teaching) wat gereeld artikels rakende leesonderrignavorsing publiseer en verslag lewer oor statistiese analises, was beoordel en geanaliseer. Alle artikels in hierdie publikasies, van 2000-2005, wat statistiese analises gebruik, is beoordeel en geanaliseer. Die data was deur middel van beskrywende statistieke geanaliseer (d.i. frekwensietellings en persentasies). Kwalitatiewe rapportering is ook gebruik.

'n Beoordeling van ses geredelik toeganklike (aanlyn) publikasies wat navorsing oor leesonderrig gepubliseer het aangetoon dat nvorsers/outeurs hoofsaaklik op statistiese-belangrikheidstoetsing staatmaak en selde of ooit oor effekgrootte/effekomvang of vertroue-interval lesings rapporteer wanneer hulle hul resultate dokumenteer.

'n Beoordeling van die literatuur toon dat nul-hipotese betekenisvolle toetsing nog altyd 'n omstrede metode van onttrekking van inligting uit eksperimentele data, en die formulering van wetenskaplike gevolgtrekkings was. Verskeie alternatiewe of aanvullings tot nul-hipotese belangrikheidstoetsing, naamlik effekgroottes, vertroue-intervalle en maganalise is voogestel.

Die volgende teoretiese stelling was vir hierdie studie geformuleer:

> Statisties-betekenisvolle toetsing in leesnavorsingstudies behoort met verslae oor effekgrootte, maganalise en vertroue-intervalle aangevul te word. Hierbenewens behoort kwantitatiewe studies wat statistieke gebruik, soos in die vorige sin vermeld, deur kwalitatiewe studies aangevul te word. Sodoende sal 'n meer omvattende beeld van leesonderrignavorsing verkry word.

Navorsing toon dat geen enkele studie ooit 'n program of praktyk as effektief vestig nie; bowendien is dit die samevloei van bewyse uit 'n verskeidenheid van studiemodelle wat uiteindelik wetenskaplik oortuigend is. Wanneer opvoeders studies en aansprake op bewyse evalueer moet hulle nie bepaal of die studie kwantitatief of kwalitatief is nie, maar eerder of dit aan die standaarde van wetenskaplike navorsing voldoen.

Die voorgestelde raamwerk wat in hierdie studie aangebied word bestaan uit drie hoofafdelings naamlik, deel een wat fokus op die studie se beskrywing van ingryping en die ewekansige-toewydingsproses, deel twee wat fokus op die studie se versameling van data en deel drie wat fokus op die studie se rapportering van resultate en meer spesifiek die statistiese rapportering van resultate.

CHAPTER 1
INTRODUCTION

1.1 Problem statement

There are few instructional tasks more important than teaching children to read. The consequences of low achievement in reading are costly both to individuals and society. Low achievement in literacy correlates with high rates of school drop-out, poverty, and underemployment (Snow, Burns, & Griffin, 1998; Wagner, 2000). The far-reaching effects of literacy achievement have heightened the interest of educators and non-educators alike in the teaching of reading. According to the National Clearinghouse for Comprehensive School Reform (2001) and the International Reading Association (2002), successful efforts to improve reading achievement emphasise identification and implementation of *evidence-based* practices that promote high rates of achievement when used in classrooms by teachers with diverse instructional styles with children who have diverse instructional needs and interests.

Being able to recognise what characterises rigorous evidence-based reading instruction is essential to choosing the right reading curriculum (i.e., what method or approach). Denton, Vaughn and Fletcher (2003:201) state that: "It will be necessary to ensure that general classroom reading instruction is of universally *high* quality and that practitioners are prepared to effectively implement validated reading interventions". When educators are not familiar with research methodologies and findings, national and provincial departments of education may find themselves implementing fads or incomplete findings. For example, a policy may require instruction in phonics but not phonemic awareness, fluency, vocabulary, and comprehension, all of which are essential components of reading instruction (National Reading Panel, 2000). Using a curriculum based on partial findings will not result in better student reading achievement. Nearly 40% of fourth grade students in the USA are unable to read at grade level (International Reading Association, 2002), while a pilot study conducted in the Potchefstroom area in the North

West Province indicates that approximately 70% of Grade three learners are unable to read at grade level (Uys & Dreyer, 2005). The results of a national evaluation of Grade 3 learners indicate that the average scores obtained by learners, across all the provinces, for reading was 39% (RSA DoE, 2001: 51). The stakes are, therefore, too high to take chances with unproven methods of reading instruction. All children are entitled to the best preparation possible.

The choice of method of instruction is very often based on empirical research studies (National Reading Panel, 2000; International Reading Association, 2002). The selection of statistical procedures is an integral part of the research process (Brantmeier, 2004). Brown (1991:569) states that statistical language studies are "legitimate investigations into phenomena in human language learning/teaching which include the use and systematic manipulation of numbers as part of their argument". Statistical significance testing is, therefore, a prominent feature of data analysis in language learning studies and also specifically, reading instruction studies (Brown, 2004; Brantmeier, 2004).

For many years, methodologists have debated what statistical significance testing means and how it should be used in the interpretation of substantive results (e.g., Carver, 1978; Cohen, 1994; Thompson, 1996; Abelson, 1997). Researchers have long placed a premium on the use of statistical significance testing. Hagan (1997:22) states that: "The logic of the [statistical test] is elegant, extraordinarily creative, and deeply embedded in our methods of statistical inference". According to Frick (1996:379), statistical significance testing is ideal for ordinal claims.

The criticism of statistical testing is, however, growing fierce. For example, Rozeboom (1997:335) argues that:

> Null hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students ... It is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism ...

According to Nunan (1991:259), many of the applied linguistic studies analyzed in a review conducted by Teleni and Baldauf (1988) can be criticized on their research designs. Nunan (1991:259) states that:

> There are also deficiencies in the manner in which they are reported. This is particularly true of experimental studies and those employing statistical analysis ... There are also studies that violate assumptions underlying the statistical procedures employed. One particular problem is the analysis of group means through t tests or ANOVA when the n size is far too small for the analysis to be valid.

Tryon (1998:796) states that:

> The fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seventy-two years of education have resulted in miniscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless substantial ...

Schmidt and Hunter (1997:37) similarly argue that: "Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution".

Criticisms of the statistical significance testing procedure are, therefore, prevalent, and occur across many scientific disciplines. This debate is not an esoteric one left for pure statisticians to resolve; applied linguistics, educational, applied psychological and other social science researchers have taken sides and argued their points cogently (Chaudron, 1988; Krantz, 1999; Svyantek & Ekeberg, 1995; Zakzanis, 1998). A recent empirical study of four disciplines on a decade-by-decade basis found an exponential increase in criticisms across disciplines of statistical testing practices (cf. Anderson et al., 1999). These criticisms are, however, far from new (cf. Boring, 1919).

Critics of statistical significance tests have made several suggestions, with the underlying theme being for researchers to examine and interpret their data

carefully and thoroughly, rather than relying solely upon p values in determining which results are important enough to examine further and report in journals. Specific suggestions include the use of effect sizes, confidence intervals, and power analyses (cf. Cohen, 1994; Kirk, 1996; Thompson, 1996; Wilkinson & The Task Force on Statistical Inference, 1999).

These criticisms and suggestions eventually led to a very important change in the 1994 American Psychological Association (APA, 1994:18) publication manual: an "encouragement" to always report effect sizes. Yet, 11 empirical studies now show that this encouragement has had no effect on the actual reporting practices within either one or two volumes of 23 journals in psychology and education (e.g., Kirk, 1996; Thompson, 1999c; Thompson & Snyder, 1998).

Indeed, the published report of the APA task force on statistical inference states that "reporting and interpreting effect sizes ... is essential to good research", and that researchers should "always present effect sizes for primary outcomes" (Wilkinson & the Task Force on Statistical Inference, 1999: 599). Yet, the Task Force (1999:599) itself acknowledged that: "Unfortunately, empirical studies of various journals indicate that the effect size of this [APA publication manual] encouragement has been negligible".

The following research questions need to be addressed:

- What is the state of affairs with regard to statistical significance testing in reading instruction research, with specific reference to post-1999 literature?
- What are the criticisms as well as the defences that have been offered for statistical significance testing?
- What are the alternatives or supplements to statistical significance testing that may be relevant for reading instruction research?

1.2 Purpose of the study

The purpose of this study was to:

- determine what the state of affairs is with regard to statistical significance testing in reading instruction research, with specific reference to post-1999 literature (post-1999 literature was selected because of the specific request, made by Wilkinson and the Task Force on Statistical Inference in 1999, to include the reporting of effect sizes in empirical research studies);
- determine what the criticisms as well as the defences are that have been offered for statistical significance testing;
- determine what the alternatives or supplements are to statistical significance testing in reading instruction research;
- To provide a framework for the most effective and appropriate selection, use and representation of statistical significance testing in the reading instruction research field.

1.3 Central theoretical statement

Statistical significance tests should be supplemented with accurate reports of effect size, power analyses and confidence intervals in reading research studies. In addition, quantitative studies, utilising statistics as stated in the previous sentence, should be supplemented with qualitative studies in order to obtain a more comprehensive picture of reading instruction research.

1.4 Method of research

1.4.1 Literature review

A critical review and analysis of the literature on statistical significance testing, as used within the reading instruction research field, was undertaken. The following databases were consulted in the Ferdinand Postma library, namely ERIC, MLA, NAVO and GKPV.

1.4.2 Empirical study

A comprehensive survey on the use of statistical significance testing, as manifested in randomly selected journals, was undertaken. Six journals (i.e., *System, Language Learning and Technology, The Reading Matrix, Scientific*

Studies of Reading, Teaching English as a Second or Foreign Language (TESL-EJ); South African Journal for Language Teaching) regularly including articles related to reading instruction research and publishing articles reporting statistical analyses, were reviewed and analysed. All articles in these journals from 2000-2005, employing statistical analyses were reviewed and analysed. The data was analysed by means of descriptive statistics (i.e., frequency counts and percentages). Qualitative reporting was also utilized.

1.5 Programme of study

Chapter 2 focuses on the making of decisions related to reading instruction research.

Chapter 3 focuses on a critical review of the statistical significance testing controversy.

Chapter 4 gives an outline of alternatives to statistical significance testing.

Chapter 5 focuses on the research methodology employed in this study.

Chapter 6 presents the results and a discussion thereof.

Chapter 7 presents a framework for selecting, using and representing statistics in reading instruction research studies.

Chapter 8 focuses on the conclusion and recommendations for future research.

CHAPTER 2

MAKING INSTRUCTIONAL DECISIONS BASED ON EVIDENCE-BASED READING INSTRUCTION RESEARCH

2.1 Introduction

Quality teachers ask tough questions in their classrooms. They want their students to probe, investigate, and inquire about what they are learning and doing. Today, teachers and administrators are also being asked tough questions about the evidence for the effectiveness of the reading programmes and approaches or methods they select for use in their classrooms (International Reading Association, 2002). The challenge in reading instruction research is to increase the efficacy of instruction in schools by identifying the instructional practices and activities that best serve to develop children's reading abilities (Snow, Burns & Griffin, 1998).

It is generally accepted that the purpose of scientific inquiry is to advance the knowledge base of humankind by seeking evidence of a phenomenon via valid experiments. In the language learning arena, specifically reading instruction, the confirmation of a phenomena should give language teachers, who teach reading, confidence in their methods, and policy makers confidence that their policies and/or curriculum will lead to better reading achievement for learners.

The purpose of this chapter is to discuss the growing demand for "evidence-based" research to guide reading instruction intervention programmes and practices. The focus is on defining evidence-based reading instruction, discussing the indicators of effectiveness, giving an outline of the methodological focus followed by the National Reading Panel (NRP), highlighting the concerns expressed about their methodological approach and discussing the future of evidence-based reading instruction research.

2.2 Evidence-based reading instruction

On April 13, 2000 a major development in the field of reading occurred with the release of the US National Reading Panel's (NRP) report titled *Teaching Children to Read*. The National Reading Panel was initiated and funded by the Congress of the United States of America. In 1997, the US Congress passed legislation which called upon a branch of the National Institutes of Health (NIH) to work with the US Department of Education in order to create a Panel to identify research-based evidence on how best to teach children to read (National Reading Panel, 2000).

Educators recognise that improving students' reading achievement is a significant challenge. All students must read proficiently. To improve every student's reading skills, policymakers must put in place carefully developed comprehensive literacy policies that include, at a minimum, reading standards and a curriculum based on findings from scientific research (NCLB, 2002).

According to the U.S. Department of Education (2002), it is necessary to create a culture and practice that demands more from educational systems by using evidence-based practices in order to transform education in the same order of magnitude as in medicine and agriculture.

2.2.1 Defining evidence-based reading instruction

The International Reading Association (2002:1) states that: "evidence-based reading instruction means that a particular programme or collection of instructional practices has a record of success. That is, there is reliable, trustworthy, and valid evidence to suggest that when the programme is used with a particular group of children, the children can be expected to make adequate gains in reading achievement". Other terms that are sometimes used to convey the same idea are research-based instruction and scientifically based research (International Reading Association, 2002).

Scientifically based research is defined in the No Child Left Behind (NCLB) legislation as "research that involves the application of rigorous, systematic,

and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs" (NCLB, 2002).

Teachers, administrators and policy makers should, therefore, be spending time and money on "what works"; instructional programmes, practices and activities that are likely to make a strong impact on learner reading achievement (Comprehensive School Reform Program Office, 2002). It is essential that especially teachers who have to implement reading instruction programmes, practices and activities are familiar with what constitutes scientific evidence of effectiveness.

2.2.2 Indicators of scientific/evidence-based effectiveness

Teachers need to take a wide range of important research into account when making instructional decisions on a day-to-day basis. More than ever, the concept of "evidence-based reading instruction" confronts reading professionals at every level. Behind the concept of evidence-based instruction lies the notion of scientifically valid and replicable research that can help teachers and administrators make effective choices.

The US NCLB (2002) (Title IX, Part A, Section 9101 [37]) legislation and the International Reading Association (2002:1) list the following indicators of effectiveness:

i) Research that employs systematic, empirical methods that draw on observation or experiment.

The defining principle of scientific evidence is systematic empiricism. Empiricism is "watching the world," relying on careful observation of events to make conclusions (Stanovich & Stanovich, 2003: 33). Systematic empiricism requires doing those observations in a careful manner in order to answer a specific question. Within the field of reading research, systematic empiricism requires an exact definition of the intervention and reading programme being studied, and a careful measurement of its outcomes.

This indicator requires quantitative research, the hallmark of which is the use of numerical measurement of student outcomes (Slavin, 2003; Brown, 2004). In order to know if one method truly caused an improvement, it is necessary to quantify the improvement in student performance.

ii) Research that involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn.

It is necessary to analyze data from a study using appropriate statistical procedures that can support the conclusions (Brantmeier, 2004). Failure to apply the appropriate statistical procedures calls the results into question. Reputable research does not issue strong claims for the effectiveness of a programme or practice based on modest differences or gains in student achievement. It is necessary to use statistics to determine whether the results were significant and important.

iii) Research that relies on measurements or observational methods that provide ***reliable*** *and* ***valid*** *data across evaluators and observers, across multiple measurements and observations, and across studies by the same or different researchers.*

Scientific research needs to use reliable methods of collecting data. A reliable testing instrument will give you the same result each time you use it on the same person or situation. Whenever a study evaluates students in a manner that relies on human judgment, as with assessments of writing ability, it is essential for the research to report *interrater reliability*, an index of how closely the different raters agree (Hatch & Lazaraton, 1991). Studies that rely on testing instruments typically establish *test-retest reliability* by administering it to the same group of people twice (Hatch & Lazaraton, 1991; Brown, 1992a, 1992b). Internal consistency (measured by Cronbach alpha) is also a type of reliability. The important point is that scientific/evidence-based research documents the reliability of its procedures for data collection.

iv) Research that is evaluated using ***experimental or quasi-experimental designs*** *in which individuals, entities, programmes, or activities are*

assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, with a preference for random-assignment experiments, or other designs to the extent that those designs contain within-condition or across-condition controls.

Experimental design. This indicator specifies that in order to be deemed scientific by the NCLB Act (2002), research needs to conform to an experimental or quasi-experimental design. The reasoning is that it is difficult to understand the effectiveness of any reading approach/method without comparing it to a different approach/method. For this reason, this indicator states that evidence for the effectiveness of any approach/method/intervention needs to include a *comparison group* to show what would happen if that practice had not been used (Seliger & Shohamy, 1989; Hatch & Lazaraton, 1991). An ideal comparison group is similar in every important way that could influence the outcome of interest. Because the comparison group allows researchers to control for the influence of external factors unrelated to the intervention, it is sometimes called a control group. By contrast, the group of people (or schools) that uses the practice under investigation is typically called the *treatment or experimental group* (Hatch & Lazaraton, 1991).

This indicator makes an additional statement about comparison groups and treatment groups: The best way to assign people to these groups is through a random process. Random assignment is the hallmark of the *experimental design*. When researchers randomly assign students (or classrooms or schools) to the experimental or control groups, any given participant in the study has an equal chance of ending up in the control group or the treatment group. The purpose of this procedure is to make sure that the two groups are as equivalent as possible in terms of the background characteristics that could influence the outcome variables. Any pre-existing differences between the comparison and the treatment group can confound the results. Random assignment eliminates, for the most part, the concern that the control group comprises people (or schools) that are fundamentally different from the treatment group.

Quasi-experimental design. Very often research in the reading field does not utilize a pure experimental design, but rather a *quasi-experimental design* (Pressley, 2003). One such approach is to select a comparison group that closely matches the control group in all relevant factors (Reyna, 2002; Slavin, 2003).

It must be noted that this indicator has generated much controversy from researchers in the reading field due to what some perceive as its exclusion of legitimate methods of scientific research such as qualitative designs and other non-experimental approaches (Purcell-Gates, 2000; Flinders, 2003; Pressley, 2003).

v) Research that ensures that experimental studies are presented in sufficient detail and clarity to allow for ***replication*** *or, at a minimum, offer the opportunity to build systematically on their findings.*

Scientific research is open to the public. A person who claims to have discovered an effective reading strategies teaching technique needs to submit evidence for its effectiveness to public scrutiny. If the results are sound, and the practice is truly effective, other people should be able to get the same results. For this reason, scientific/evidence-based research must be reported in sufficient detail to allow for *replication* of the intervention and the scientific findings (Pressley, 2003).

One type of replication involves practitioners reproducing the reading instruction intervention in their own schools. Another type of replication is more demanding; it involves another researcher attempting to replicate the original findings by following the same research procedures. This is an important process because it allows researchers to independently confirm the legitimacy of purported scientific evidence. For this reason, scientific research also needs to include all of the details about the educational intervention, participants, materials, outcome measures (e.g., tests and questionnaires), and the statistical procedures that were employed (Seliger & Shohamy, 1989; Hatch & Lazaraton, 1991).

vi) Research that has been accepted by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective, and scientific review.

The process of peer review is essential to scientific/evidence-based research (International Reading Association, 2002). Many journals focusing on the publishing of reading research results, such as *TESOL Quarterly*, *Scientific Studies of Reading*, *Reading Research Quarterly*, *Journal of research in Reading*, and *Reading in a Foreign Language*, accept their articles based on the review of other researchers who understand the research topic. The purpose of peer review is to submit research to public criticism. This process helps to screen out poor quality research, especially research that has serious problems in any of the areas mentioned above. A variety of journals, with varying degrees of stringency of standards, exist, so peer review is a minimal standard. Yet because it is minimal, its absence is a sure sign that a particular method is lacking in quality (Stanovich & Stanovich, 2003). It is possible to determine whether a journal is peer reviewed by reading its editorial policy for acceptance of manuscripts.

These indicators are intended to encourage researchers to provide better and more useful evidence of what works and to challenge practitioners to make good decisions based on evidence (Feuer, 2002).

2.2.3 Research approaches excluded from the indicators

The indicators discussed in the previous section define a standard for scientific/evidence-based research that focuses on the question of "what works" in reading instruction practice (International Reading Association, 2002). There are, however, other important questions in reading instruction research for which different research approaches (e.g., qualitative and ethnographic studies) are appropriate (Purcell-Gates, 2000). Some of these questions include "How do social interactions among students influence their learning to read?" and "What different demands do learners face when reading in different disciplines, such as science, math, or social studies?" Both quantitative and qualitative research provide us with important lenses through which to study and examine critical reading instruction questions

and issues. Ethnographic research, case studies, and other qualitative designs provide a highly detailed narrative description of educational settings and reading interventions. The South African National Centre for Educational Statistics has commissioned many large-scale surveys over the past 30 years. Descriptive studies use these survey data to describe important trends in education, such as the national rate of progress in improving literacy and mathematics proficiency. Correlational studies identify the factors that correlate with important outcomes. These studies help us to understand the conditions of families, schools, and communities that promote success in reading.

Being able to recognize what characterizes rigorous research is essential to choosing the right reading programmes and/or approaches/methods and activities. The common goal for researchers and teachers alike is to make instructional choices and decisions that give all students, no matter their circumstances, the best possible chance of reaching a high level of reading achievement. To do this it is necessary to pay attention to multiple sources of valid and reliable research findings drawn from multiple research paradigms (Farstrup, 2003).

2.3 The International Reading Panel's methodological focus

Hall (n.d.: 2) states that: "The field of reading has been notorious for its lax standards of research, an issue that often is raised to explain the pendulum swings in approaches used to teach reading".

The National Reading Panel (2000) decided to select only experimental and quasi-experimental studies that were constructed to compare reading performance between groups that received a specific kind of reading instruction versus a control group. They wanted to be able to make causal statements that a particular type of instruction leads to higher reading achievement. Therefore, they generally did not include research studies that were only qualitative and descriptive without measuring outcomes of

instruction groups versus control groups (National Institute of Child Health and Human Development, 2000).

The evidence-based methodological standards adopted by the National Reading Panel are essentially those normally used in research studies of the efficacy of interventions in psychological and medical research. These include behaviourally based interventions, medications, or medical procedures proposed for use in the fostering of robust health and psychological development and the prevention or treatment of disease (NICHD, 2000). It is the view of the National Reading Panel (2000) that the efficacy of materials and methodologies used in the teaching of reading and in the prevention or treatment of reading disabilities should be tested no less rigorously.

One of the most significant outcomes of the National Reading Panel's work is that the methodology they used suggests a set of guidelines to define what high quality scientific research is in the field of reading instruction. Slavin (2003:12) states that, "At long last, education reform may be entering an era of well-researched programs and practices".

2.3.1 Criteria and coding used for study selection and analysis

The National Reading Panel determined that there were approximately 100,000 studies on reading that had been published since 1966. Five topic subgroups (alphabetics – phonemic awareness and phonics – fluency, comprehension, teacher education, and computer technology) were formed and they each searched a minimum of two research databases (typically PsychINFO and ERIC) to identify a large pool of study citations on their topic. The number of years searched varied by each subgroup so that they would identify 300-400 potential sources from the most recent years. Each citation was evaluated based on the following criteria:

- Focus of study must be on children's reading development (preschool through grade 12).
- Study must be published in English in a refereed journal.

- Used an experimental or quasi-experimental design with a control group or a multiple-baseline method (NRP, 2000: 1-5 – 1-6).

Those studies meeting the above criteria formed the set of studies subjected to further analysis. The subgroup member evaluated the study against a second set of review criteria, which were:

- Study participants must be thoroughly described (i.e., age, demographics, cognitive, academic and behavioural characteristics).
- Study interventions must be described in sufficient detail to allow for replication, including how long the interventions lasted and how long the effects lasted.
- Study methods must allow judgments about how instruction fidelity was ensured.
- Studies must include a full description of outcome measures (NRP, 2000: 1-5 – 1-6).

For all the studies that met the initial screening and the four criteria above, a subgroup member thoroughly reviewed the full research study and coded the research so that the data could be further analyzed. The coding included a substantial amount of information about the study, including:

- Research question – the question that the study addresses.
- Sample of students participating - states or countries, number of different schools, number of different classrooms, number of participants, age, grade, reading levels of participants, location (urban, rural, suburban), list of pre-tests administered, SES, ethnicity, exceptional learning characteristics (reading disabled, etc.), selection criteria, description of classroom curriculum, and attrition during the study.
- Setting of the study - classroom, laboratory, or tutorial.
- Design of the study - random assignment of participants to treatments with or without a pre-test, treatment components administered in a fixed vs. variable order, baseline description.

- Description of treatment and control conditions - nature and components of reading instruction provided to control group, implicit or explicit instruction, difficulty level of any text used, duration of treatment (minutes per session, sessions per week, and number of weeks), number of trainers, teacher/student ratio, type of trainer (classroom teacher, student teacher, researcher, clinician, special education teacher, parent, peer, other), qualifications of trainers, length and type of training given to trainers.
- Names of standardized or investigator-constructed tests to measure reading outcomes.
- Non-equivalence of groups - assessment of subgroup member on any concerns about equivalence of treatment and control groups.
- Result for each measure - name of measure, difference between treatment and control group is positive or negative and size, number of people providing the effect size information.
- Name of person coding and amount of time to complete coding (NRP, 2000: 1-7 – 1-9).

For each study meeting the above criteria, relevant reported statistics were coded in a standardized format and analysed using meta-analyses and the calculation of effect sizes. The NRP (2002) presented their major findings in the areas of alphabetics, fluency, comprehension, teacher education, and computer technology and reading instruction.

2.3.2 Concerns levelled against the NRP's methodological focus

Several researchers (Purcell-Gates, 2000; Flinders, 2003; Pressley, 2003) have expressed their concern about the NRP's exclusive focus on experimental and quasi-experimental studies to study the effectiveness of reading instruction intervention programmes/approaches/methods. Purcell-gates (2000) states that, "by limiting the type of research used to address the issue of how to guide policy and practice in reading instruction to experimental and quasi-experimental studies, the panel missed critical areas that have been examined by qualitative and ethnographic research". Similarly, Pressley (2003) states that "I think evidence-based reading

instruction is a good thing, although I am concerned about how narrowly it is being construed in many policy conversations".

Many issues and problems related to reading instruction require research that draws on multiple perspectives, approaches and procedures. Purcell-Gates (2000) states that:

> We simply do not know enough about the complexities of the complex process of learning in schools to design effective experimental studies that will solve all of our problems. We need methodologies that will allow us to probe for insights, for possible operative factors, for new information, before we can begin to think about limiting our research studies to experimental, hypothesis-testing approaches.

An additional aspect that is a cause for concern in the NRP's methodological focus is the lack of reference to studies specifically reporting the use of effect sizes (cf. NRP, 2000). Although the five subgroups of the NRP (2000) used meta-analyses and effect sizes in the analysis and reporting of their results, the specific reporting of effect sizes within the studies was not one of the criteria required for studies to be selected for further analysis.

Brown (2004:372) states that: "One of the issues that I have consistently worried about is the poor quality of much of the quantitative/statistical research that I read in the journals of our field". The selection of appropriate statistical procedures driven by research questions is a critical part of the L2 reading research process (Brantmeier, 2004).

When educators have to determine which type of reading instruction to implement it is essential that they know how big an effect the instruction under investigation had (Pressley, 2003). Educators should be particularly careful in interpreting claims that instructional effects are significant. The word significant has ambiguities associated with it when used to describe a research outcome. It can be used to state that the finding is statistically significant (e.g., there is less than a 5% or 1% or one tenth of 1% chance

that the difference obtained is a chance difference). In this sense, a finding is considered to be highly significant when there is an exceptionally low probability that the difference observed was due to chance. Thus, when there is less than one tenth of 1% chance that a difference is due to chance, the finding is often referred to as highly significant – even though such a statistical difference can be quite small in absolute or practical terms, so small as to affect performance hardly at all. Most often, when a small practical or absolute effect has high statistical significance, it is because there were many participants in each of the instructional conditions. When that is the case, a difference of high statistical significance can be inconsequential in terms of educational or practical classroom significance. That is, the students receiving the new intervention do only a little better than the students receiving the conventional instruction. There are many who are willing to praise interventions they favour because they produce effects of high statistical significance that are really very small absolute, practical, or educational effects. Educators show demand to know just how big statistically significant effects are in terms of practical or absolute effect sizes. The effect size is a description of how large an effect the treatment had. It should be reported in real-world terms, such as percentage of children reading at or above grade level (Coalition for Evidence-Based Policy, 2003). The size of an effect indicates its importance. Some effects can be statistically significant, but of such a small magnitude that they are unimportant.

The statistical significance debate and the reporting, or not, of effect sizes is discussed in detail in chapters 3 and 4.

2.4 The future of scientific evidence-based reading instruction research

Educators have long given lip service to research as a guide to practice (Slavin, 2003). Increasingly, they are being asked to justify their choices of programmes and practices using the findings of rigorous, experimental research. According to Slavin (2003), scientific research traditionally has played a relatively minor role in educational reform since many innovative

practices and programmes are untested. When reform efforts fail, educators and policymakers move to implement a different set of innovations that also have untested claims, instead of adopting well-researched programmes and practices that have been proven to work. Shifting to a new paradigm will mean changing practice to look more deeply to research-based programmes rather than following a new trend.

Grossen (1996:22) states that:

> Unlike most other fields of scientific inquiry, education places extraordinary emphasis on the new and the novel. Believing that the most recent theory – at whatever level of research – is also the most important, education leaders may lose sight of the value of seminal research and proven practices.

Survey research conducted in South Africa indicates that learners from primary through tertiary level are experiencing problems with their reading ability (Dreyer, 1998; RSA Department of Education, 2001; Van Wyk, 2001). It seems appropriate and essential that the South African education system also accept, use and implement scientific evidence-based research as part of its educational initiatives. It will require building a data-driven system in which educators look for results and continually challenge themselves to make decisions for improvement based on the integration of meaningful and relevant data and research.

It will require educators to let go of past practices that are no longer justifiable because they are not achieving identified outcomes. It will require risk taking to understand and apply research methods in classrooms so that educators will not only be guided by evidence but also informed of their own results. It will require a new vision of pre- and in-service language teacher development that seeks to empower educators to be the agents of change and to try new ideas that offer to improve their learners' reading proficiency and reading abilities. The impetus to move forward must come from the willingness to use scientific evidence-based research and apply it to classroom settings. After all, learners deserve no less than access to that which is proven effective and that which offers hope that improvements in

reading achievement can be made. Slavin (2003: 16) states that, "Evidence-based reform could finally bring education to the point reached early in the 20th century by medicine, agriculture, and technology, fields in which evidence is the lifeblood of progress". Proper consideration of scientific evidence-based research gives educators greater confidence in their decision making and may lead to greater opportunity for students to succeed.

2.5 Conclusion

The quest to find the "best programmes" for teaching reading has a long and quite unsuccessful history (International Reading Association, 2002). Despite many claims of programme excellence, literacy scholars (e.g., Allington, 2001; Stahl et al., 1998) argue that careful examination of such studies reveals the use of either flawed designs or selective reporting of the available data.

Quantitative studies generally investigate programme effects on relatively large numbers of students. In contrast, qualitative studies typically focus on small samples or on individuals and are especially valuable in helping teachers understand how particular reading programmes or approaches affect individuals who may not represent the mainstream or average student.

However, a review of the literature seems to indicate that no single study ever establishes a programme or practice as effective; moreover it is the convergence of evidence from a variety of study designs that is ultimately scientifically convincing. When evaluating studies any claims of evidence, educators must not determine whether the study is quantitative or qualitative in nature, but rather if the study meets the standards of scientific research. That is, does it involve "rigorous and systematic empirical inquiry that is data-based" (Bogdan & Biklen, 1992:43).

The challenge that confronts teachers and administrators is the need to view the evidence that they read through the lens of their particular school and classroom settings. They must determine if the instructional strategies

and routines that are central to the materials under review are a good match for the particular learners they teach.

In addition to examining the match between the instructional approach or programme and the learners they teach, teachers must also consider the match between the instructional approach or programme and the resources available for implementation.

Pearson (quoted in The National Right to Read Foundation, 2003) commented on the need for teacher educators to prepare teachers who are knowledgeable about evidence-based reading instruction. "Somehow we must generate the political will to insist that every teacher receive quality training on all aspects of research-based pedagogy".

Each and every teacher education programme should ensure that its students know how to read reports of the very latest research so that they can change their classroom practices to match new evidence.

CHAPTER 3
STATISTICAL SIGNIFICANCE TESTING

3.1 Introduction

Researchers have been conducting statistical testing for research purposes since the early 1700s (McLean & Ernest, 1998; Huberty, 1993). In the history of science many researchers have resorted to tests of significance when testing hypotheses. The legacies of Sir Ronald Fisher (concerning differences of between and within groups using probability levels) and Karl Pearson (concerning correlation analyses providing indices of association) are two important approaches of statistical testing and how statistical analyses have developed (McLean & Ernest, 1998; Nix & Barnette, 1998).

The role of statistical significance testing in research methodology literature has been the subject of much controversy (Kaufman, 1998; Knapp, 1998; Levin, 1998; McLean & Ernest, 1998; Nix & Barnette, 1998; Thompson, 1998). As early as 1931, R. W. Tyler noted the misuse of statistical significance, ". . . we are prone to conceive of statistical significance as equivalent to social significance. These two terms are essentially different and ought not to be confused" (quoted in Daniel, 1998:24). Berkson (1942), Yates (1951), Kerlinger (1979) and Daniel (1998) also lamented that too much emphasis was placed upon statistical significance tests as the end all product. Thompson (1998) states that, "the weakest link in contemporary quantitative educational research involves the methodologies of statistical analysis". In essence, the controversy involves the sole use (and misinterpretation) of the p-value without taking into account other descriptive statistics that provide a broader picture of the data analysis.

The purpose of this chapter is to give an overview of the controversy regarding statistical significance testing by focusing on the debate (i.e., presenting arguments for and against statistical significance testing) as well as analysing the key concepts involved in statistical significance testing, namely the role of chance and Type I and Type II errors.

3.2 An overview of the statistical significance testing debate

Educational and specifically reading researchers continually strive for rigorous, objective approaches in making valid inferences concerning scientific questions in these disciplines. The dominant, traditional approach has been to frame the question in terms of two contrasting statistical hypotheses: one representing no difference between population parameters of interest (i.e., the null hypothesis, H_0) and the other representing either a unidirectional or bidirectional alternative (i.e., the alternative hypothesis, H_a) (Hatch & Lazaraton, 1991).

3.2.1 Fisher versus Neyman and Pearson

Nix and Barnette (1998:4) state that:

> Most who read recent textbooks devoted to statistical methods are inclined to believe statistical significance testing is a unified, non-controversial theory whereby we seek to reject the null hypothesis in order to provide evidence of the viability of the alternative hypothesis.

This is, however, not true. Fisher proposed the testing of a single binary null hypothesis using the p-value as the strength of the statistics. He did not develop or support the alternative hypotheses, type I and type II errors in significance testing, or the concept of statistical power. Therefore, according to Fisher's p-value approach, after stating a null hypothesis and then obtaining sample results, the probability of obtaining the sample result, or a more extreme result, is computed, assuming that the null hypothesis is true in the population from which the sample was derived (Cohen, 1994; Thompson, 1996).

In contrast to Fisher's notion of null hypothesis significance testing (NHST), Pearson and Neyman viewed significance testing as "a method of selecting a hypothesis from a slate of candidate hypotheses, rather than testing of a single hypothesis" (Nix & Barnette, 1998:4). The Neyman-Pearson or fixed-alpha approach specifies a level at which the test statistic should be rejected and is set a priori to conducting the test of data. A null hypothesis (H_0) and an alternative hypothesis (H_a) are stated, and if the value of the

test statistic falls in the rejection region, the null hypothesis is rejected in favour of the alternative hypothesis. Otherwise the null hypothesis is retained on the basis that there is insufficient evidence to reject it.

Distinguishing between the two methods of statistical testing is important in terms of how methods of statistical analysis have developed in the recent past. Fisher's legacy of statistical analysis approaches (including ANOVA methods) relies on subjective judgments concerning differences between and within groups, using probability levels to determine which results are statistically significant from each other. Pearson's legacy involves the development of correlational analyses and providing indexes of association. It is because of different approaches to analyses and different philosophical beliefs that the issue of testing for statistical significance has arisen.

Currently, we are in an era where the value of statistical significance testing is being challenged by many researchers (Cohen, 1994; Nester, 1996; Thompson, 1998; Anderson et al., 2000; Kline, 2004). It seems, however, that very few reading instruction researchers are aware of the debate regarding statistical significance testing among statisticians. The Language Learning field has lagged behind other disciplines with respect to awareness and discussion of problems associated with statistical significance testing. Both positions (arguing for and against the use of statistical significance tests in research) are presented in the next two sections.

3.2.2 Arguing for statistical significance testing

A perusal of some of the most popular journals in language learning and specifically those journals publishing research on reading instruction indicate that statistical significance testing has seemingly withstood all of the criticisms, as it remains a widely-used analytical tool in this field (Brantmeier, 2004). The controversy is based on whether or not statistical significance testing has value in answering a research question posed by researchers (Anderson et al., 2000). This section addresses several reasons why statistical significance testing has weathered the storm, and why researchers still use statistical significance tests. The reasons covered include: a) the usefulness of statistical significance testing in making

categorical statements and testing ordinal claims, b) researchers' dissatisfaction with the alternatives to statistical testing, and c) the argument that statistical testing as originally conceived is a logical and sound method of statistical analysis, and persistent misuse is the fault of the researchers, rather than an indication of inherent flaws within the method (Frick, 1996; Abelson, 1997; Harris, 1997).

3.2.2.1 Usefulness in testing ordinal claims

Ordinal claims are defined as those that do not specify size of effect; they specify only order or direction. Thus, "A is larger than B", and "reading strategy use is positively correlated with reading comprehension achievement", are examples of ordinal claims because they provide no information about effect size or strength of association. Frick (1996: 379) noted that "for quantitative claims, null hypothesis testing is not sufficient ..., but for ordinal claims it is ideal". By an ordinal claim, Frick (1996:380) meant a claim that does not specify the size of an effect but that specifies "only the order of conditions, the order of effects, or the directions of a correlation". Tukey (1991) contends that it is really the direction of an effect rather than the existence of one that the t-test helps decide. Others have argued that if the question of interest is whether a difference is positive or negative, significance testing is a suitable approach to finding out (Abelson, 1997; Harris, 1997).

According to Abelson (1997), Frick (1996), and Greenwald et al. (1996), the goal of science is not always determining size of effect; testing ordinal claims (i.e., directional hypotheses) and making categorical statements (i.e., asserting that something is important) are also important goals of science, goals for which statistical significance testing is well-suited.

3.2.2.2 Lack of superior alternatives

Researchers in favour of statistical significance testing state that the proposed alternative methods, such as effect sizes and confidence intervals, are less informative than statistical significance tests, and are equally vulnerable to widespread misinterpretation (Frick, 1996; Harris, 1997). For example, Harris (1997:10) stated that statistical significance

testing "provides useful information that is not easily gleaned from the corresponding confidence interval: degree of confidence that we have not made a Type III error and likelihood that our sample result is replicable".

Dixon (1998:391) contended that despite justified claims of the limitations of significance testing and misinterpretations of test results, p values can convey useful information regarding the strength of evidence: "Although p values are not the most direct index of this information, they provide a reasonable surrogate within the constraints posed by the mechanics of traditional hypothesis testing".

Cortina and Dunlap (1997) concluded that statistical significance tests and proposed alternatives such as confidence intervals each have something valuable to contribute to science, and therefore should be used in conjunction with each other.

3.2.2.3 Misuse does not equal inherent flaws

Supporters of statistical significance tests argue that these methods are not inherently flawed; rather, years of misuse of this logical and potentially useful tool have gradually led to its disrepute (Abelson, 1997; Cortina & Dunlap, 1997; Frick, 1996; Hagen, 1997). Hagen (1997:22) states that:

> The logic of the [statistical test] is elegant, extraordinarily creative, and deeply embedded in our methods of statistical inference. It is unlikely that we will ever be able to divorce ourselves from that logic even if someday we decide that we want to ... The [statistical test] has been misinterpreted and misused for decades. This is our fault, not the fault of the [statistical test] ... The logic underlying statistical significance testing has not yet been successfully challenged.

Similarly, Hubbard et al. (1997:550) state that:

> From the researcher's (and possibly journal editor's and reviewer's) perspective, the use of significance tests offers the prospect of effortless, cut-and-dried decision-making concerning the viability of a hypothesis. The role of informed judgment and intimate familiarity with the data is largely superseded by rules of thumb with

dichotomous, accept-reject outcomes. Decisions based on tests of significance certainly make life easier.

According to Cortina and Dunlap (1997), careful judgment is required in all areas of science, including statistical analysis, and that the "cure" for misuse and misinterpretation lies not in banning the method, but in improving our education and refining our judgment. Indeed, "mindless application of any procedure because it has been misapplied ensures the proverbial loss of both baby and bathwater" (Cortina & Dunlap, 1997: 171).

3.2.3 Arguing against statistical significance testing

Several important issues have fuelled the arguments against the use of statistical significance tests. A review of the post-1994 literature indicates that the most often cited issues include: a) replication, b) sample size, c) what statistical significance tests actually tell us, and d) practical significance.

3.2.3.1 Replication

One of the most powerful arguments against the use of statistical significance testing is that these analyses tell neither the researcher nor the research consumer anything about the replication of a study's results. According to Thompson (1994:157), the importance of replication in research has enjoyed increased awareness as:

> Social scientists have increasingly recognized that the single study is inherently governed by subjective passion, that ideology frequently drives even analytic choices, and that the projection against potentially negative consequences of these passions occurs not from feigned objectivity, but arises in the aggregate across studies from an emphasis on replication.

The increased role of replication in educational research has been accompanied by a growing realization that statistical significance testing has severely limited utility, especially with regard to evaluating the likely replication of study results (Cohen, 1994; Greenwald et al., 1996; Thompson, 1994; 1995).

If the purpose of science is formulating generalisable insight based on the accumulation of findings that will generalise under stated conditions, and if the most promising strategies to fulfil this purpose emphasise interpretation based on the estimated likelihood that results will replicate, then statistical significance tests are rendered virtually useless for the underlying purpose of science. Thompson (1994, 1995) has proposed and elaborated upon the use of several methods that researchers can employ to empirically assess the internal replication of their research results; these methods include cross-validation, the bootstrap, and the jackknife.

The reason that statistical significance tests do not evaluate result replication is that, notwithstanding common misperceptions to the contrary (Cohen, 1994), statistical significance tests do not test the probability that sample results occur in the population (Carver, 1978). As Thompson (1996:27) explained:

> Put succinctly, $P_{CALCULATED}$ is the probability (0 to 1.0) of the sample statistics, given the sample size, and assuming the sample was derived from a population in which the null hypothesis (H_0) is exactly true.

Statistical significance tests, therefore, assume (not test) the population, and test (not assume) the sample results. As Cohen (1994) explained, this is not what researchers want to do. But as he also noted, the statistical significance test "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (Cohen, 1994:997).

3.2.3.2 Sample size

Although the likelihood that a true null hypothesis will be rejected does not increase with the sizes of the samples compared, the likelihood that a real difference of a given magnitude will result in rejection of the null hypothesis at a given level of confidence does (Thompson, 1996; Zakzanis, 1998). It is also the case that the smaller a real difference is, the larger the samples are likely to have to be to provide a basis for rejecting the null hypothesis. In

other words, whether or not one assumes that the null hypothesis is always or almost always false, when it is false the probability that a statistical significance test will lead to rejection increases with sample size (Asraf & Brewer, 2004; Nickerson, 2000; Thompson, 1998).

It means that conclusions drawn from experiments often depend on decisions researchers have made regarding how many participants to include in the study. Also, inasmuch as even very small real differences will be detected by sufficiently large samples, it is possible with very large samples to demonstrate statistical significance for differences that are too small to be of any theoretical or practical interest (Thompson, 1996; 1998; Nickerson, 2000). Thus, as Hays (1981:293) argued almost 25 years ago, "virtually any study can be made to show significant results if one uses enough subjects". And as Thompson (1995:85) explained:

> Because statistical significance tests largely evaluate the size of the researcher's sample, and because researchers already know prior to conducting significance tests whether the sample in hand was large or small, outcomes of these statistical tests do not always yield new insight as a return for the effort invested in conducting the tests.

A decision to either reject or not reject the null hypothesis is, therefore, largely dependent upon the researcher's sample size. Thompson (1998:799) states that, "Statistical testing becomes a tautological search for enough participants to achieve statistical significance. If we fail to reject, it is only because we've been lazy to drag in enough participants".

3.2.3.3 What statistical significance tests actually tell us

Many researchers feel that an overemphasis on statistical significance testing detracts researchers from the primary purposes and goals of science, such as interpreting research outcomes, theory development, and formulating generalisable insight based on the accumulation of scientific findings (Kirk, 1996; Schmidt, 1996; Thompson, 1995). Kirk (1996: 753-754) states that "even when a significance test is interpreted correctly, the business of science does not progress as it should ... What we want to know is the size of the difference between A and B and the error associated with our estimate; knowing that A is greater than B is not enough".

Researchers are, therefore, of the opinion that while statistical significance tests may be useful in determining the direction of relationships, it is also necessary to know the strength or magnitude of relationships or differences, and statistical tests are not effective in this regard.

3.2.3.4 Practical significance

Many researchers are concerned with the ubiquitous practice of equating statistically significant findings with findings that are of practical importance. That is, many researchers present their data in such a way that findings that are found to be statistically significant are also interpreted to be useful, meaningful, or important. Kirk (1996: 746) defined the difference between statistically significance and practical significance as: "Statistical significance is concerned with whether a research result is due to chance or sampling variability; practical significance is concerned with whether the result is useful in the real world".

In addition, Cohen (1994: 1001) stated that:

> ... statistically significant does not mean plain-English significant, but if one reads the literature, one often discovers that a finding reported in the Results section studded with asterisks implicitly becomes in the Discussion section highly significant or very highly significant, important, big!

According to some researchers (Cohen, 1990, 1994; Kirk, 1996; Thompson, 1998), because we usually know in advance that the null hypothesis is false, the rejection of a null hypothesis is not very informative or important. What are important are measures of the strength of association between the independent and dependent variables and measure of effect size (Cohen, 1994; Kirk, 1996; Thompson & Snyder, 1998; Thompson, 1996, 1999a, 1999b, 1999c).

3.3 Statistical significance testing: Focussing on the concepts

In this section the purpose is to analyse the key concepts involved in the statistical significance testing controversy.

3.3.1 Assessing the role of chance

There are two basic statistical methods used to assess the role of chance: hypothesis testing and confidence intervals (cf. Figure 1).

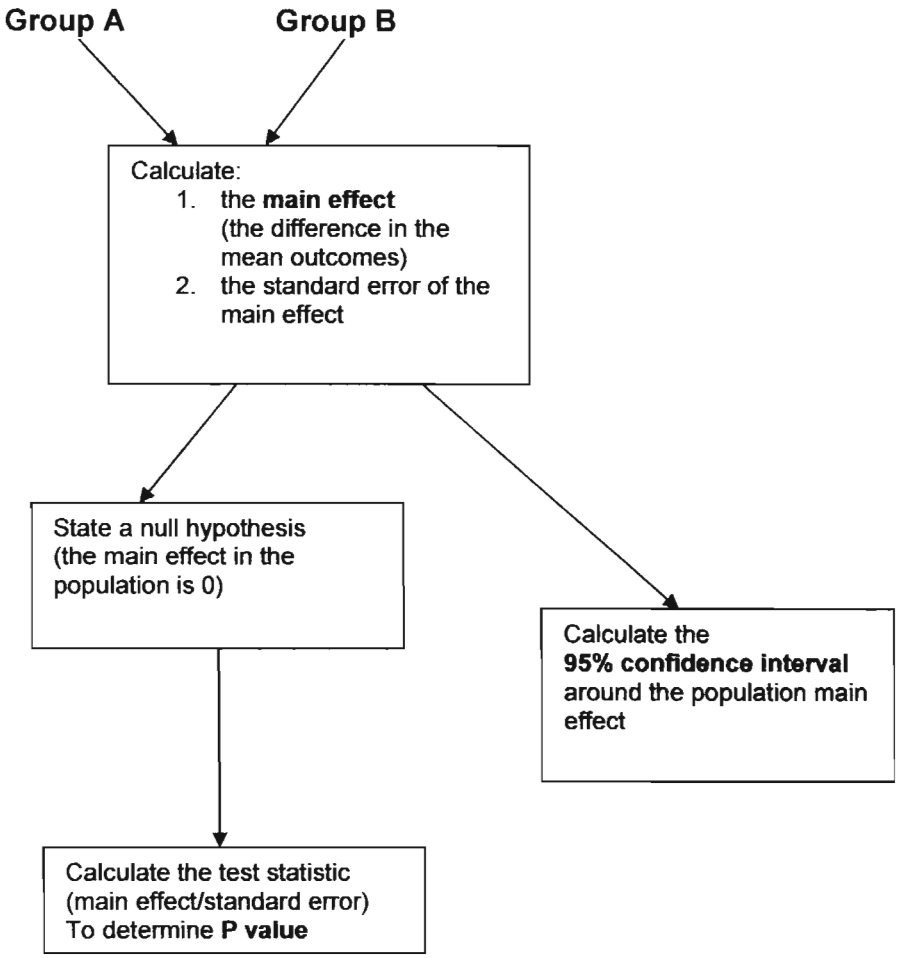


Figure 1: Statistical approach to comparing two groups

3.3.1.1 Hypothesis testing

Scientific study is frequently based around the concept of testing hypotheses. A hypothesis is a theory or statement of belief about the population of interest (Hatch & Lazaraton, 1991) (e.g., that there is a difference in the mean reading strategy use of English Second Language (ESL) learners in Grades 4-7 in School A and in School B). In order to see how likely it is that the theory is true, data of samples from Schools A and B is used to test what is called the null hypothesis, the complement of the theory of interest (Hatch & Lazaraton, 1991). So, instead of trying to show that the mean reading strategy use of ESL learners in School A and School B differs, researchers determine whether they have enough evidence in the sample to disprove the null hypothesis that there is no difference in reading strategy use in the populations of ESL learners in Grades 4-7 in schools A and B. However, there will always be a degree of uncertainty associated with the inferences that can be drawn. If researchers had examined a different sample of ESL learners from School A and School B they might have come to a different conclusion.

The null hypothesis (H_0), therefore, implies that there is no difference in the two population means. Researchers such as Bakan (1966), Cohen (1988), and Hinkle et al. (1994) have called this difference, the hypothesis of no difference. According to Carver (1978:381):

> The null hypothesis states that the experimental group and the control group are not different with respect to [a specified property of interest] and that any difference found between their means is due to sampling fluctuation.

Further development of Fisher's null hypothesis resulted in the null hypothesis indicating direction (e.g., $\mu_1 \leq \mu_2$ or $\mu_1 \geq \mu_2$). Conversely, the research question or the alternative hypothesis (H_a) indicates that there is a difference between two population means (e.g., $\mu_1 \neq \mu_2$) and this hypothesis may also be directional (e.g., $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$) (Portillo, 2001).

The way in which researchers decide whether or not they have enough evidence to reject the null hypothesis is by considering what is called the P-

value (Hatch & Lazaraton, 1991:231). This is a probability and so lies between 0 and 1. An event cannot occur if P=0 and it must occur if P=1. The P-value represents the probability of getting observed results (or more extreme results) if the null hypothesis is true. A small P-value indicates that the results in the sample would be unlikely to arise if the null hypothesis were true; this implies that the null hypothesis is not true, and there is evidence to reject it. Usually a cut-off value of P=0.05 is selected, so that any value of P<0.05 is called "small" and leads to rejection of the null hypothesis. If the null hypothesis is rejected, the result is described as statistically significant at that level (Nickerson, 2000:243). According to Glaser (1999), misinterpretations of the P value and what it entails are not unusual. Part of this misinterpretation may stem from the lack of uniformity in the definition of the P value. Even though the exact terminology may differ (cf. Pedhazur & Schmelkin, 1991; Moore, 1995; Thompson, 1996), one generally agreed upon definition of the P value is that the null hypothesis is ultimately being tested against a level of significance designated by the researcher.

According to most textbooks, null hypothesis testing admits of only two possible decision outcomes: rejection (at a specified significance level) of the hypothesis of no difference, and failure to reject this hypothesis (at that level) (Seliger & Shohamy, 1989; Hatch & Lazaraton, 1991; Brown & Rodgers, 2002). Given the latter outcome, one is justified in saying only that a significant difference was not found; one does not have a basis for concluding that the null hypothesis is true (that the samples were drawn from the same population with respect to the variable of interest) (Nickerson, 2000). Inasmuch as the null hypothesis may be either true or false and it may either be rejected or fail to be rejected, any given instance of null hypothesis significance testing admits of four possible outcomes (cf. Table 1). There are two ways to be right: rejecting the null hypothesis when it is false (when the samples were drawn from different populations) and failing to reject it when it is true (when the samples were drawn from the same population). There are also two ways to be wrong: rejecting the null hypothesis when it is true (Type I error) and failing to reject it when it is false (Type II error) (cf. section 3.3.2).

Consider a study of a reading strategies intervention programme: Group A receives the intervention and improves their reading comprehension scores on average by 10 percent, while group B serves as a control and improves their reading comprehension scores by an average of 3 percent. The main effect of the reading strategies intervention programme is therefore estimated to be a 7 percent increase (i.e., 10-3) in reading comprehension scores (on average). But we would rarely expect that any two groups would have exactly the same increase in their reading comprehension scores. So could it just be chance that group A improved their reading comprehension scores by 10 percent?

Hypothesis testing, therefore, considers the null hypothesis that no difference exists. In the above-mentioned example, the null hypothesis is that the reading comprehension score change in the two groups is the same. The test addresses the question, "If the true state of affairs is no difference (i.e., the null hypothesis is true), what is the probability of observing this difference (i.e., 7%) or one more extreme (i.e., 8%, 9%, etc.)"? This probability is called the *P* value and translates roughly to "the probability that the observed result is due to chance."

If the *P* value is less than 0.05 (5%), researchers typically assert that the findings are "statistically significant." In the case of the reading strategies intervention programme, if the chance of observing a difference of 7% or more (when, in fact, none exists) is less than 5 in 100, then the reading strategies intervention programme is presumed to have a real effect.

3.3.1.1.1 Factors that influence p values

Statistical significance (meaning a low *P* value) depends on three factors: the main effect itself and the two factors that make up the standard error. Here is how each relates to the *P* value (American College of Physicians, 2001:183):

- *The magnitude of the main effect*. A 7% difference will have a lower *P* value (i.e., more likely to be statistically significant) than a 1% difference.
- *The number of observations*. A 7% difference observed in a study with 500 participants in each group will have a lower *P* value than a 7% difference observed in a study with 25 participants in each group.
- *The spread in the data (commonly measured as a standard deviation)*. If everybody in group A increases their reading comprehension scores by approximately 10% and everybody in group B increases their reading comprehension scores by approximately 3%, the *P* value will be lower than if there is a wide variation in individual reading comprehension score changes (even if the group averages remain at 10 and 3 percent). More observations do not reduce spread in data.

3.3.1.2 Confidence intervals

Researchers and readers frequently face questions about the role of chance in a study's results. The traditional approach has been to consider the probability that an observed result is due to chance, the P value. However, P values provide no information on the results' precision, that is, the degree to which they would vary if measured multiple times. Consequently, an increasing emphasis is being placed on a second approach: reporting a range of plausible results, better known as the 95% confidence interval (CI). This section focuses on the concept of confidence intervals and their relationship to p values.

Consider the reading strategies intervention programme example mentioned in section 3.3.1.1. Group A receives the intervention and increases their reading comprehension scores by 10%, whereas group B serves as a control and increases their reading comprehension scores by an average of 3%. The main effect of the reading strategies intervention programme is therefore estimated to be a 7% increase in reading comprehension scores (on average).

But readers should recognize that the true effect of the programme may not be exactly a 7% reading comprehension score increase. Instead, the true effect is best represented as a range. What is the range of effects that might be expected with high probability? If the probability is taken as, for example, 0.95, CI is called a 95% CI. That is the question addressed by a CI. For example:

The mean reading comprehension score increase was 10% for participants in the intervention group and 3% for participants in the control group, resulting in a mean difference of 7% and a 95% CI of 2 to 12. In other words, 95% of the time the true effect (i.e., the main effect of the populations from which the samples were drawn) of the intervention will be within the range from 2 to 12 percent.

To conceptualize the more formal definition of a 95% CI, it is useful to consider what would happen if the study were repeated 100 times. Obviously, not every study would result in a 7% reading comprehension score increase in favour of the intervention. Simply due to the play of chance, reading comprehension score increase would be greater in some studies and less in others, and some studies might show that the controls increased their reading comprehension scores by more than that of the intervention group. One would then expect 95 out of the 100 differences to be within the interval 2 to 12 percent.

3.3.1.2.1 Factors that influence 95% confidence intervals

Confidence intervals are a measure of how precise an estimated effect is. The range of a CI is dependent on the two factors that cause the main effect to vary (American College of Physicians, 2001: 229-230):

1) *The number of observations*. This factor is largely under the researcher's control. A 7% difference observed in a study with 500 participants in each group will have a narrower CI than a 7% difference observed in a study with 25 participants in each group.

2) *The spread in the data* (commonly measured as a standard deviation). This factor is largely outside the researcher's control. Consider the two

comparisons in Figure 2. In both cases, the reading comprehension gain in group A is 10% and the reading comprehension gain in group B is 3%. If everybody in group A improves near 10% and everybody in group B improves near 3%, then the CI will be narrower (left part of figure) than if individual reading comprehension gains are spread all over the map (right part of figure).

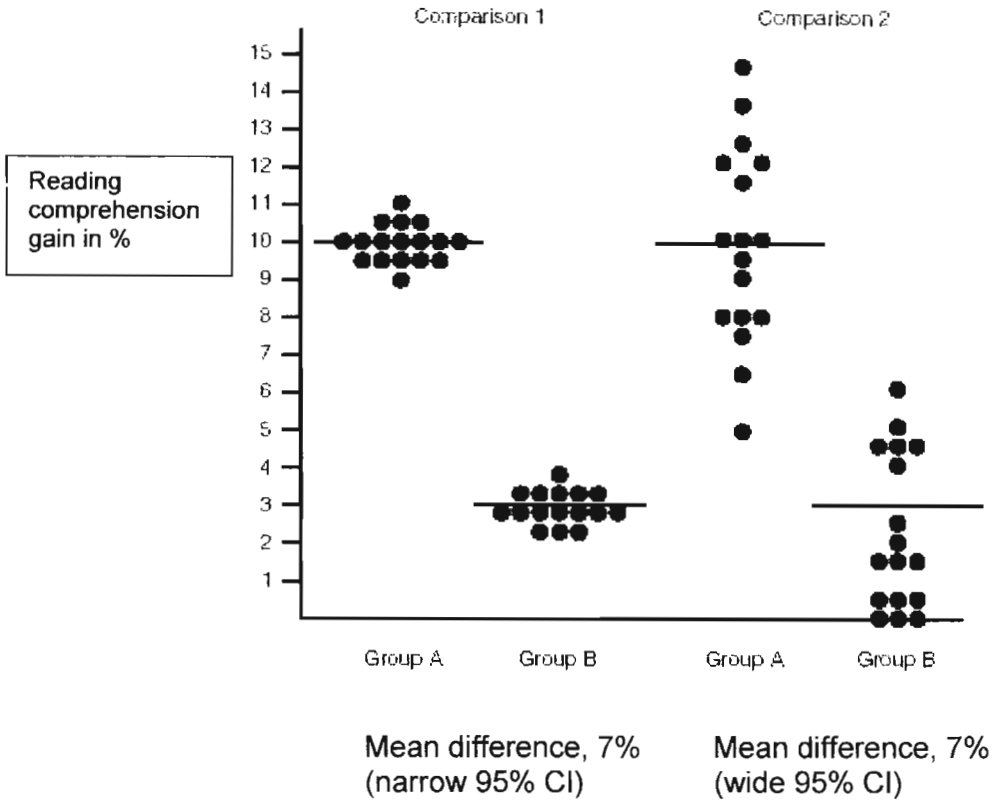


Figure 2: More diversity in reading comprehension scores equals a larger 95% CI. The horizontal bars represent group means

(Adapted from: American College of Physicians, 2001:230)

3.3.1.2.2 Relationship between 95% confidence intervals and p values

Information about the *P* value is contained in the 95% CI. As shown in Figure 3, the *P* value can be inferred based on whether the finding of "no difference" falls within the CI.

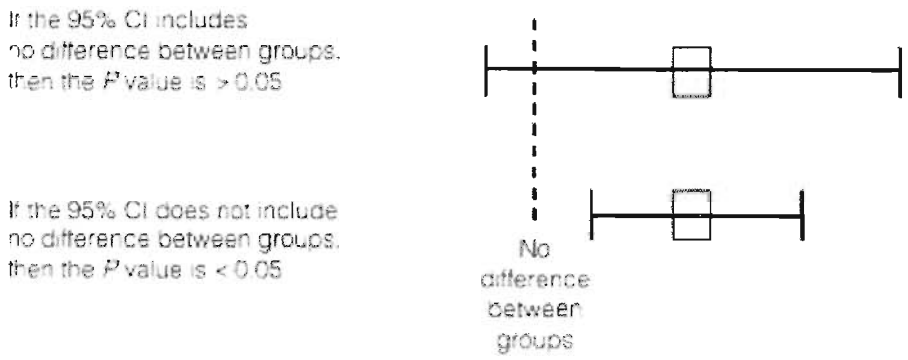


Figure 3: Relationship between P value and 95% CI

(Adapted from: American College of Physicians, 2001:230)

So, given a CI of 2 to 12 percent for the true difference, one could infer that the *P* value is less than 0.05. Alternatively, given a CI of -3 to 17 percent for the true difference, one could infer that the *P* value is greater than 0.05. If the CI terminates exactly on no difference, such as 0 to 14 percent, then the *P* value is exactly 0.05.

Although *P* values and 95% CIs are related, CIs are preferred because they convey information about the range of plausible effects. In other words, the CI provides the reader with some sense of how precise the estimate of the effect is. This is a valuable dimension that is not contained within a *P* value.

3.3.2 Type I and Type II errors

Statistical tests are tools that help us assess the role of chance as an explanation of patterns observed in data. The most common "pattern" of interest is how two groups compare in terms of a single outcome. After a statistical test is performed, researchers (and readers) can arrive at one of two conclusions:

- The pattern is probably not due to chance (i.e., "There was a significant difference" or "The study was positive").
- The pattern is likely due to chance (i.e., "There was no significant difference" or "The study was negative").

No matter how well the study is performed, either conclusion may be wrong. A mistake about the first conclusion is labeled a Type I error and a mistake about the second is labeled a Type II error (Asraf & Brewer, 2004) (cf. Table 1).

Table 1: Results of Null Hypothesis Statistical Testing

	Truth state of H_0	
Decision regarding H_0	**False**	**True**
Rejected	Correct rejection	Type I error
Not rejected	Type II error	Correct nonrejection

(Nickerson, 2000: 243).

3.3.2.1 Type I Errors

A Type I error involves rejecting the null hypothesis when the null hypothesis is in fact true (a false positive, for example, the treatment was effective when it was not) (Asraf & Brewer, 2004). The probability of making a Type I error is equal to the value of the researcher's selected alpha (α) level (Portillo, 2001: 209). If the researcher chooses an alpha level equal to 0.05, the probability of committing a Type I error is five times out of one hundred. Therefore, as the researcher lowers the alpha level, the probability of committing a Type I error is lowered. However, as the researcher lowers the probability of committing a Type I error, the researcher then sacrifices the power of the test (Portillo, 2001). The probability that a type I error has occurred in a positive study is the exact *P* value reported. For example, if the *P* value is 0.001, then the probability that the study has yielded false-positive results is 1 in 1000.

3.3.2.2 Type II Errors

A type II error is analogous to a false-negative result during diagnostic testing: No difference is shown when in "truth" there is one (Asraf & Brewer, 2004). The power of the test (1-β) "is the probability that a test statistic will find statistical significance" (Rossi, 1997:177 cited in Nix & Barnette, 1998). A test with power of 0.80 indicates the researcher would have an 80 percent

chance of finding statistical significance. Since power is defined as 1-β, beta (β) represents Type II error (Portillo, 2001). When the researcher accepts the null hypothesis and the null hypothesis is false, a false negative results (e.g., no treatment effect present when there was) (Asraf & Brewer, 2004). Traditionally, this error has received less attention from researchers than type I error and, consequently, may occur more often. Type II errors are generally the result of a researcher studying too few participants. To avoid the error, some researchers perform a sample size calculation before beginning a study and, as part of the calculation, assert what a "true difference" is and accept that they will miss it 10% to 20% of the time (i.e., Type II error rate of 0.1 or 0.2).

3.4 Conclusion

Null hypothesis significance testing has been and is a controversial method of extracting information from experimental data and of guiding the formation of scientific conclusions. Meehl (1997:421) states that:

> Competent scholars persist in strong disagreement, ranging from some who think H_0-testing is pretty much all right as practice, to others who think it is never appropriate. Most critics fall somewhere between these extremes, and they differ among themselves as to their main *reasons* for complaint.

One hypothesis worth entertaining about the controversy is that it is a tempest in a teapot (Nickerson, 2000). A minimal goal for research within reading instruction should be to attempt to achieve a better understanding among researchers of the approach, of its strengths and limitations, of the various objections that have been raised against it, and of the assumptions that are necessary to justify specific conclusions that can be drawn from its results.

Statistical methods should facilitate good thinking, and only to the degree that they do so are they being used well. When applied unthinkingly in cookbook fashion and without awareness of their limitations and of the assumptions that are needed to justify conclusions drawn from them, they

can get in the way of productive reasoning; when used judiciously, with cognizance of their limitations, they can be very helpful in making sense of data and determining what conclusions are justified.

CHAPTER 4
ALTERNATIVE ANALYSES

4.1 Introduction

Despite years of criticism, null hypothesis significance testing continues to be widely used by researchers in various disciplines. Although it is important to point out the difficulties with hypothesis testing (cf. Chapter 3), it is also necessary to contemplate alternatives that could either complement or replace null hypothesis significance testing. Given the growing demand for evidence-based research to guide reading instruction interventions, interest in reporting research results using alternatives to significance testing is growing (Boston, 2003-2004).

The purpose of this chapter is to determine what feasible alternatives to null hypothesis significance testing are available for use by reading instruction researchers. Each of the alternatives, namely effect size, confidence intervals, and power analysis are discussed and evaluated for their ability to either replace or complement null hypothesis significance testing.

4.2 Effect size

Perhaps the most commonly recommended alternative to solely interpreting P values is to determine the magnitude of effect, more commonly known as "effect size" (Kotrlik & Williams, 2003) The routine use of effect sizes, however, has generally been limited to meta-analysis, for combining and comparing estimates from different studies, and is all too rare in original reports of educational research (Keselman et al., 1998). This is despite the fact that measures of effect size have been available for at least 60 years (Huberty, 2002), and the American Psychological Association has been officially encouraging authors to report effect sizes since 1994, but with limited success (Wilkinson & APA Task Force on Statistical Inference, 1999). Prior to considering effect size measures as alternatives to hypothesis testing, a brief review of their nature is in order.

4.2.1 Defining effect size and its importance

The term effect size has become increasingly popular throughout educational and reading instruction literature in recent years. The APA Task Force emphasized that researchers should "always provide some effect-size estimate when reporting a P value ... reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (Wilkinson & APA Task Force on Statistical Inference, 1999: 599). With these discussions focusing on effect size, the literature presents a variety of effect size definitions. Common definitions include: a standardized value that estimates the magnitude of the differences between groups (Thomas, Salazar & Landers, 1991), a standardized mean difference (Olejnik & Algina, 2000; Vacha-Haase, 2001), the degree to which sample results diverge from the null hypothesis (Cohen, 1988; 1994), a measure of the difference or association deemed large enough to be of "practical significance" (Morse, 1998), an estimate of the degree to which the phenomenon being studied exists in the population (e.g., a correlation or difference in means) (Hair et al., 1995), and strength of relationship (American Psychological Association, 2001).

Among the numerous effect size definitions, the majority of definitions include the descriptors "standardized difference between means" or "standardized measure of association". In practice, researchers commonly use two categories of measures of effect size in the literature, namely measures of effect size (according to group mean differences), and measures of strength of association (according to variance accounted for) (Maxwell & Delaney, 1990).

An early discussion of effect size by Karl Pearson (1901) addressed the idea that statistical significance provides the reader with only part of the story and therefore must be supplemented. Whereas statistical tests of significance tell us the likelihood that experimental results differ from chance expectations, effect size measurements tell us the relative magnitude of the experimental treatment. They tell us the size of the experimental effect. Effect sizes are especially important because they allow us to compare the magnitude of experimental treatments from one experiment to another. Although percent

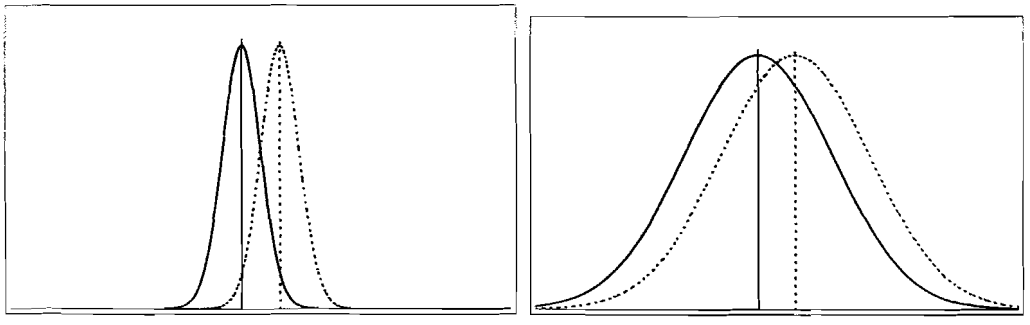
improvements can be used to compare experimental treatments to control treatments, such calculations are often difficult to interpret and are almost always impossible to use in fair comparisons across experimental paradigms (Thalheimer & Cook, 2002). Reporting effect size allows a researcher to judge the magnitude of the differences present between groups, increasing the capability of the researcher to compare current research results to previous research and to judge the practical significance of the results derived. Support for reporting effect size is witnessed throughout the literature. For example, Fan (2001) described good research as presenting both statistical significance testing results as well as effect sizes. Baugh (2002:255) declared, "Effect size reporting is increasingly recognized as a necessary and responsible practice".

Consider an experiment conducted by Dowson (2000) to investigate time of day effects on learning: do children learn better in the morning or afternoon? A group of 38 children were included in the experiment. Half were randomly allocated to listen to a story and answer questions about it (on tape) at 9 am, the other half to hear exactly the same story and answer the same questions at 3 pm. Their comprehension was measured by the number of questions answered correctly out of 20.

The average score was 15.2 for the morning group, 17.9 for the afternoon group: a difference of 2.7. The question may now be asked as to how big a difference is this? If the outcome were measured on a familiar scale, such as GCSE grades, interpreting the difference would not be a problem. If the average difference were, say, half a grade, most people would have a fair idea of the educational significance of the effect of reading a story at different times of day. However, in many experiments there is no familiar scale available on which to record the outcomes. The experimenter often has to invent a scale or to use (or adapt) an already existing one – but generally not one whose interpretation will be familiar to most people.

One way to get over this problem is to use the amount of variation in scores to contextualise the difference (Coe, 2002). If there were no overlap at all and every single person in the afternoon group had done better on the test than

everyone in the morning group, then this would seem like a very substantial difference. On the other hand, if the spread of scores were large and the overlap much bigger than the difference between the groups, then the effect might seem less significant. Because we have an idea of the amount of variation found within a group, we can use this as a yardstick against which to compare the difference. This idea is quantified in the calculation of the *effect size*. The concept is illustrated in Figure 4, which shows two possible ways the difference might vary in relation to the overlap in the distribution of scores (displayed as bell-shaped frequency curves). If the difference were as in graph (a) it would be very significant; in graph (b), on the other hand, the difference might hardly be noticeable.



(a) (b)

Figure 4: Two possible ways the difference might vary in relation to the overlap

(Coe, 2002).

4.2.2 Calculating effect size

Although extensive articles have been written detailing methods for calculating effect sizes from published research articles (e.g., Rosnow & Rosenthal, 1996; Rosnow et al., 2000), at least for the majority of practicing reading instruction teachers, these calculations might be too technical. Cohen's ***d*** has two advantages over other effect size measurements. First, its burgeoning popularity is making it the standard. Thus, its calculation enables immediate comparison to increasingly larger numbers of published studies. Second, Cohen's (1992) suggestion that effect sizes of 0.20 are small, 0.50 are medium, and 0.80 are large enables one to compare an experiment's effect size results to known benchmarks (Thalheimer & Cook, 2002).

In essence, an effect size is the difference between two means (e.g., treatment minus control) divided by the standard deviation of the two conditions. In other words:

$$\textbf{Effect Size} = \frac{\textbf{[Mean of experimental group] – [Mean of control group]}}{\textbf{Standard Deviation}}$$

It is the division by the standard deviation that enables one to compare effect sizes across experiments. The 'standard deviation' is a measure of the spread of a set of values. It can refer either to the standard deviation of the control group, or to a 'pooled' value from both groups (Thalheimer & Cook, 2002:4), assuming equal standard deviations.

4.2.3 Interpreting effect size

One feature of an effect size is that it can be directly converted into statements about the overlap between the two samples in terms of a comparison of percentiles (Coe, 2002).

An effect size is exactly equivalent to a 'Z-score' of a standard Normal distribution. For example, an effect size of 0.8 means that the score of the average person in the experimental group is 0.8 standard deviations above the average person in the control group, and hence exceeds the scores of 79% of the control group. With the two groups of 19 in the time-of-day effects experiment (Dowson, 2000), the average (median) person in the 'afternoon' group (i.e., the one who would have been ranked 10^{th} in the group) would have scored about the same as the 4^{th} highest person in the 'morning' group. Visualising these two individuals can give quite a graphic interpretation of the difference between the two effects.

Table 2 shows conversions of effect sizes (column 1) to percentiles (column 2) and the equivalent change in rank order for a group of 25 (column 3). For example, for an effect-size of 0.6, the value of 73% indicates that the average person in the experimental group would score higher than 73% of a control group that was initially equivalent. If the group consisted of 25 people, this is the same as saying that the average person (i.e. ranked 13^{th} in the group) would now be on a par with the person ranked 7^{th} in the control group. An effect-size of 1.6 would raise the average person to be level with the top ranked individual in the control group, so effect sizes larger than this are illustrated in terms of the top person in a larger group. For example, an effect size of 3.0 would bring the average person in a group of 740 level with the previously top person in the group.

Table 2: Interpretations of effect sizes

Effect Size	Percentage of control group who would be below average person in experimental group	Rank of person in a control group of 25 who would be equivalent to the average person in experimental group	Probability that you could guess which group a person was in from knowledge of their 'score'.	Equivalent correlation, *r* (=Difference in percentage 'successful' in each of the two groups, BESD)	Probability that person from experimental group will be higher than person from control, if both chosen at random (=CLES)
0.0	50%	13th	0.50	0.00	0.50
0.1	54%	12th	0.52	0.05	0.53
0.2	58%	11th	0.54	0.10	0.56
0.3	62%	10th	0.56	0.15	0.58
0.4	66%	9th	0.58	0.20	0.61
0.5	69%	8th	0.60	0.24	0.64
0.6	73%	7th	0.62	0.29	0.66
0.7	76%	6th	0.64	0.33	0.69
0.8	79%	6th	0.66	0.37	0.71
0.9	82%	5th	0.67	0.41	0.74
1.0	84%	4th	0.69	0.45	0.76
1.2	88%	3rd	0.73	0.51	0.80
1.4	92%	2nd	0.76	0.57	0.84
1.6	95%	1st	0.79	0.62	0.87
1.8	96%	1st	0.82	0.67	0.90
2.0	98%	1st (or 1st out of 44)	0.84	0.71	0.92
2.5	99%	1st	0.89	0.78	0.96

		(or 1st out of 160)			
3.0	99.9%	1st (or 1st out of 740)	0.93	0.83	0.98

(Coe, 2002).

Another way to conceptualise the overlap is in terms of the probability, normal distributions should therefore be assumed, that one could guess which group a person came from, based only on their test score - or whatever value was being compared. If the effect size were 0 (i.e. the two groups were the same) then the probability of a correct guess would be exactly a half - or 0.50. With a difference between the two groups equivalent to an effect size of 0.3, there is still plenty of overlap, and the probability of correctly identifying the groups rises only slightly to 0.56. With an effect size of 1, the probability is now 0.69, just over a two-thirds chance. These probabilities are shown in the fourth column of Table 2. It is clear that the overlap between experimental and control groups is substantial (and therefore the probability is still close to 0.5), even when the effect-size is quite large.

A slightly different way to interpret effect sizes makes use of an equivalence between the standardised mean difference (*d*) and the correlation coefficient, *r*. If group membership is coded with a dummy variable (e.g. denoting the control group by 0 and the experimental group by 1) and the correlation between this variable and the outcome measure calculated, a value of *r* can be derived. By making some additional assumptions, one can readily convert *d* into *r* in general, using the equation $r^2=d^2/(4+d^2)$ (see Cohen, 1969: 20-22 for other formulae and conversion table). Rosenthal and Rubin (1982) take advantage of an interesting property of *r* to suggest a further interpretation, which they call the binomial effect size display (BESD). If the outcome measure is reduced to a simple dichotomy (for example, whether a score is above or below a particular value such as the median, which could be thought of as 'success' or 'failure'), *r* can be interpreted as the difference in the proportions in each category. For example, an effect size of 0.2 indicates a

difference of 0.10 in these proportions, as would be the case if 45% of the control group and 55% of the treatment group had reached some threshold of 'success'. However, if the overall proportion 'successful' is not close to 50%, this interpretation can be somewhat misleading (Strahan, 1991; McGraw, 1991). The values for the BESD are shown in column 5.

Finally, McGraw and Wong (1992) have suggested a 'Common Language Effect Size' (CLES) statistic, which they argue is readily understood by non-statisticians (shown in column 6 of Table 2). This is the probability (normality assumed) that a score sampled at random from one distribution will be greater than a score sampled from another. They give the example of the heights of young adult males and females, which differ by an effect size of about 2, and translate this difference to a CLES of 0.92. In other words "n 92 out of 100 blind dates among young adults, the male will be taller than the female" (McGraw & Wong, 1992:361).

4.2.4 Effect size versus statistical significance

Effect size quantifies the size of the difference between two groups, and may therefore be said to be a true measure of the significance of the difference (Coe, 2002; Kotrlik & Williams, 2003). If, for example, the results of Dowson's (2000) 'time of day effects' experiment were found to apply generally, we might ask the question: 'How much difference would it make to children's learning if they were taught a particular topic in the afternoon instead of the morning?' The best answer we could give to this would be in terms of the effect size.

However, in statistics the word 'significance' is often used to mean 'statistical significance', which is the likelihood that the difference between the two groups could just be an accident of sampling. If you take two samples from the same population there will always be a difference between them. The statistical significance is usually calculated as a 'P-value', the probability that a difference of at least the same size would have arisen by chance, even if

there really were no difference between the two populations. For differences between the means of two groups, this P-value would normally be calculated from a statistical procedure known as a 't-test'. By convention, if $P < 0.05$ (i.e., below 5%), the difference is taken to be large enough to be 'significant'; if not, then it is 'not significant' (Hatch & Lazaraton, 1991).

According to Coe (2002) there are a number of problems with using 'significance tests' in this way. The main one is that the P-value depends essentially on two things: the size of the effect *and* the size of the sample (Carver, 1993; Thompson, 1996; 1998). One would get a 'significant' result either if the effect were very big (despite having only a small sample) or if the sample were very big (even if the actual effect size were tiny). It is important to know the statistical significance of a result, since without it there is a danger of drawing firm conclusions from studies where the sample is too small to justify such confidence. However, statistical significance does *not* tell you the most important thing: *the size of the effect.* One way to overcome this confusion is to report the effect size, together with an estimate of its likely 'margin for error' or 'confidence interval' (Thompson, 1996; 1998).

The confusion between effect size and statistical significance seems to stem from misconceptions of what statistical significance testing tells the researcher (Denis, 2003). Nickerson (2000) explains that, there is a belief that a small value of P means a treatment effect of large magnitude, and that statistical significance means theoretical or practical significance. Olejnik and Algina (2000:241) state that: "Statistical significance testing does not imply meaningfulness". Statistical significance testing evaluates the probability of obtaining the sampling outcome by chance, while effect size provides some indication of practical meaningfulness (Fan, 2001). Statistical significance relies heavily on sample size, while effect size assists in the interpretation of results and makes trivial effects harder to ignore, further assisting researchers to decide whether results are practically significant (Kirk, 2001).

An example in a study demonstrates why the reporting of effect size is important. In this example, failing to report effect size would have resulted in the researcher using the results of a statistically significant t-test to conclude that important differences existed. In this study, Williams (2003) compared the percent of time that faculty actually spent in teaching with the percent of time they preferred to spend in teaching. The data show that even though the t-test was statistically significant (t=2.20, P=0.03, df=154), Cohen's effect size value (d=0.09) did not meet the standard of even a "small" effect size. This indicated that the difference has low practical significance, so Williams made no substantive recommendations based on the results of this t-test.

4.2.5 Factors influencing effect size

Although effect size is a simple and readily interpreted measure of effectiveness, it can also be sensitive to a number of spurious influences, so some care needs to be taken in its use (Coe, 2002).

4.2.5.1 Which standard deviation?

The first problem is the issue of which 'standard deviation' to use. Ideally, the control group will provide the best estimate of standard deviation, since it consists of a representative group of the population who have not been affected by the experimental intervention. However, unless the control group is very large, the estimate of the 'true' population standard deviation derived from only the control group is likely to be appreciably less accurate than an estimate derived from both the control and experimental groups. Moreover, in studies where there is not a true 'control' group then it may be an arbitrary decision which group's standard deviation to use, and it will often make an appreciable difference to the estimate of effect size (Thalheimer & Cook, 2002).

For these reasons, it is often better to use a 'pooled' estimate of standard deviation. The pooled estimate is essentially an average of the standard deviations of the experimental and control groups (Equation below). This is not the same as the standard deviation of all the values in both groups

'pooled' together. If, for example each group had a low standard deviation but the two means were substantially different, the true pooled estimate (as calculated by the equation below) would be much lower than the value obtained by pooling all the values together and calculating the standard deviation. The implications of choices about which standard deviation to use are discussed by Olejnik and Algina (2000).

$$SD_{pooled} = \sqrt{\frac{(N_E - 1)SD_E^2 + (N_C - 1)SD_C^2}{N_E + N_C - 2}}$$

(Where N_E and N_C are the numbers in the experimental and control groups, respectively, and SD_E and SD_C are their standard deviations.)

The use of a pooled estimate of standard deviation depends on the assumption that the two calculated standard deviations are estimates of *the same* population value. In other words, that the experimental and control group standard deviations differ only as a result of sampling variation. Where this assumption cannot be made (either because there is some reason to believe that the two standard deviations are likely to be systematically different, or if the actual measured values are very different), then a pooled estimate should not be used (Coe, 2002). The maximum of the two SDs gives a 'conservative' estimate of the effect size (Steyn, 2000:1-3).

4.2.5.2 Non-normal distributions

The interpretations of effect sizes given in Table 2 depend on the assumption that both control and experimental groups have a 'Normal' distribution (i.e., the familiar 'bell-shaped' curve, shown, for example, in Figure 4). If this assumption is not true then the interpretation regarding overlap and probabilities, but not the effect size itself, may be altered, and in particular, it may be difficult to make a fair comparison between an effect size based on Normal distributions and one based on non-Normal distributions.

An illustration of this is given in Figure 5, which shows the frequency curves for two distributions, one of them Normal, the other a 'contaminated normal'

distribution (Wilcox, 1998), which is similar in shape, but with somewhat fatter extremes. In fact, the latter does look just a little more spread-out than the Normal distribution, but its standard deviation is actually over three times as big. The consequence of this in terms of effect-size differences is shown in Figure 6. Both graphs show distributions that differ by an effect-size equal to 1, but the appearance of the effect-size difference from the graphs is rather dissimilar. In graph (b), the separation between experimental and control groups seems much larger, yet the effect-size is actually the same as for the Normal distributions plotted in graph (a). In terms of the amount of overlap, in graph (b) 97% of the 'experimental' group are above the control group mean, compared with the value of 84% for the Normal distribution of graph (a) (as given in Table 2). This is quite a substantial difference and illustrates the danger of using the values in Table 2 when the distribution is not known to be Normal (Coe, 2002).

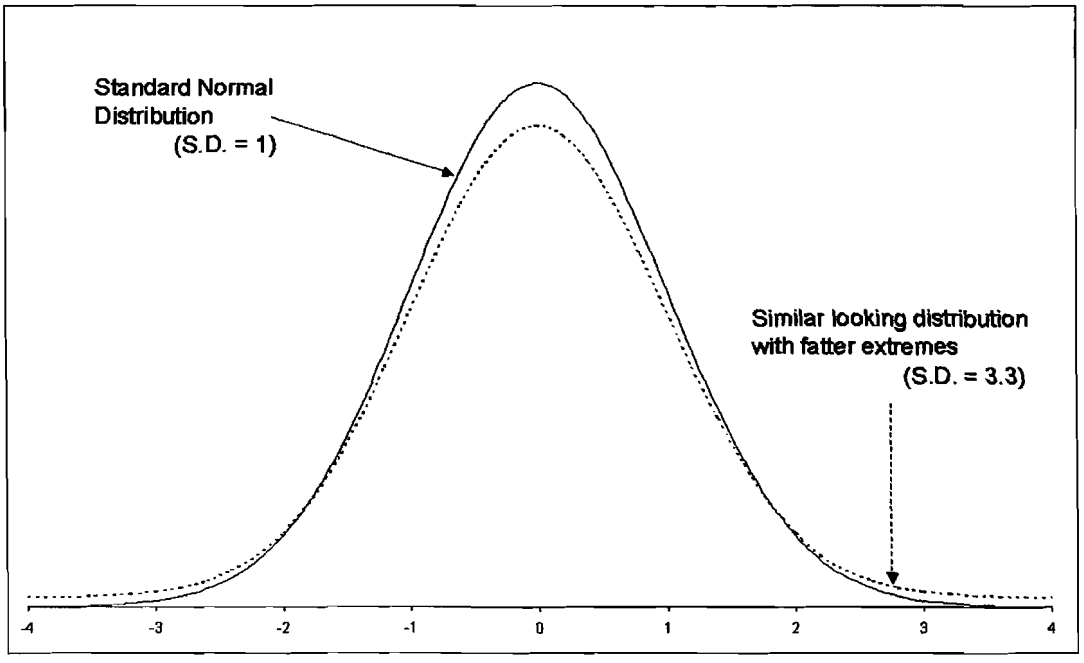
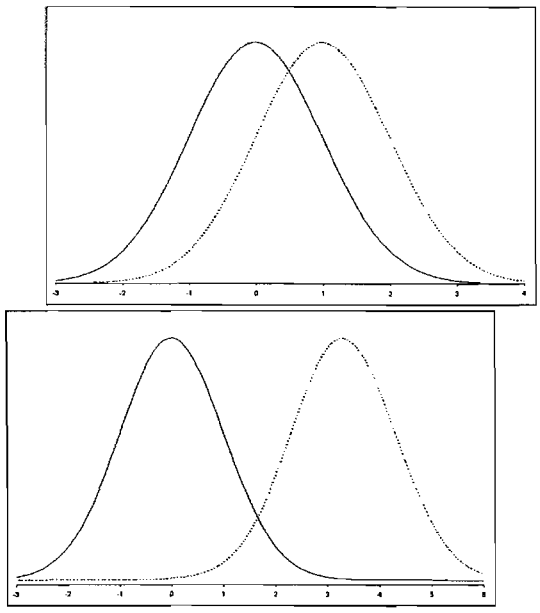


Figure 5: Comparison of Normal and non-Normal distributions

(Coe, 2002).



(a) (b)

Figure 6: Normal and non-Normal distributions with effect-size = 1

(Coe, 2002).

4.2.5.3 Measurement reliability

An additional factor that can affect an effect-size is the reliability of the measurement on which it is based. According to classical measurement theory, any measure of a particular outcome may be considered to consist of the 'true' underlying value, together with a component of 'error'. The problem is that the amount of variation in measured scores for a particular sample (i.e., its standard deviation) will depend on both the variation in underlying scores and the amount of error in their measurement (Thompson, 1998; Baugh, 2002).

To give an example, imagine giving a reading comprehension test at different times of the day (morning and afternoon). The experiment was conducted twice with two (hypothetically) identical samples of students. In the first version the test used to assess their reading comprehension consisted of just 10 items and their scores were converted into a percentage. In the second version a test with 50 items was used, and again converted to a percentage.

The two tests were of equal difficulty and the actual effect of the difference in time-of-day was the same in each case, so the respective mean percentages of the morning and afternoon groups were the same for both versions. However, it is almost always the case that a longer test will be more reliable, and hence the standard deviation of the percentages on the 50 item test will be lower than the standard deviation for the 10 item test. Thus, although the true effect was the same, the calculated effect sizes will be different.

In interpreting an effect size, it is therefore important to know the reliability of the measurement from which it was calculated. This is one reason why the reliability of any outcome measure used should be reported. It is theoretically possible to make a correction for unreliability (sometimes called 'attenuation'), which gives an estimate of what the effect size would have been, had the reliability of the test been perfect (Hatch & Lazaraton, 1991; Baugh, 2002). However, in practice the effect of this is rather alarming, since the worse the test was, the more you increase the estimate of the effect size. Moreover, estimates of reliability are dependent on the particular population in which the test was used, and are themselves subject to sampling error (Baugh, 2002).

4.2.6 Alternative measures of effect size

A number of statistics are sometimes proposed as alternative measures of effect size, other than the 'standardised mean difference'. Some of these are briefly referred to.

4.2.6.1 Parametric correlations

Perhaps the simplest measures and reporting of effect sizes exist for correlations. The correlation coefficient itself is a measure of effect size. The most commonly used statistics for parametric correlations are Pearson's *r*, Spearman's *rho*, and point-biserial (Hatch & Lazaraton, 1991; Hopkins, 1997). The size of these correlation coefficients must be interpreted using a set of descriptors for correlation coefficients as described in Table 3.

Table 3: Descriptors for interpreting effect size

Source	Statistic	Value	Interpretation
Rea & Parker, 1992	Phi or Cramér's V	.00 and under .01	Negligible association
		.10 and under .20	Weak association
		.20 and under .40	Moderate association
		.40 and under .60	Relatively strong association
		.60 and under .80	Strong association
		.80 and under 1.00	Very strong association
Cohen, 1988	Cohen's d	.20	Small effect size
		.50	Medium effect size
		.80	Large effect size
	Cohen's f	.10	Small effect size
		.25	Medium effect size
		.40	Large effect size
Cohen, 1988	R^2	.0196	Small effect size
		.1300	Medium effect size
		.2600	Large effect size
Kirk, 1996	Omega squared	.010	Small effect size
		.059	Medium effect size
		.138	Large effect size
Davis, 1971[a]	Correlation coefficients	.70 or higher	Very strong association
		.50 to .69	Substantial association
		.30 to .49	Moderate association
		.10 to .29	Low association
		.01 to .09	Negligible association
Hinkle, Wiersma, & Jurs, 1979[a]	Correlation coefficients	.90 to 1.00	Very high correlation
		.70 to .90	High correlation
		.50 to .70	Moderate correlation
		.30 to .50	Low correlation
		.00 to .30	Little if any correlation
Hopkins (1997)[a]	Correlation coefficients	.90 to 1.00	Nearly, practically, or almost: perfect, distinct, infinite
		.70 to .90	Very large, very high, huge
		.50 to .70	Large, high, major
		.30 to .50	Moderate, medium
		.10 to .30	Small, low, minor
		.00 to .10	Trivial, very small, insubstantial, tiny, practically zero

(Kotrlik & Williams, 2003: 5).

4.2.6.2 Simple and multiple regression

Another simple measure of effect size is the square of the multiple regression coefficient, also known as the coefficient of determination, R^2 (Cohen, 1988).

All standard statistical analysis packages calculate this coefficient automatically, and this coefficient represents the proportion of variance in the dependent variable explained by the independent variable(s) through a linear multiple regression. The effect size for this coefficient may be interpreted using a set of descriptors derived from Cohen's f^2 statistic (see Table 3).

4.2.6.3 Calculating Cohen's d from t-tests

Researchers should list the means of the treatment condition and the comparison condition. The researchers should also list the number of subjects and the standard deviations of the treatment condition and the comparison condition. Use those numbers in the formula and calculate the pooled standard deviation using the formula below.

$$SD_{pooled} = \sqrt{\frac{(N_E - 1)SD_E^2 + (N_C - 1)SD_C^2}{N_E + N_C - 2}}$$

(Where N_E and N_C are the numbers in the experimental and control groups, respectively, and SD_E and SD_C are their standard deviations.)

After doing the above calculation the formula below should be used to complete the calculation.

$$d = \frac{\overline{x}_t - \overline{x}_c}{s_{pooled}}$$

Key to symbols:

d = Cohen's d effect size

$\overline{x}$ = mean (average of treatment or comparison conditions)

s = standard deviation

Subscripts: t refers to the treatment condition and c refers to the comparison condition (or control condition).

(Rosnow & Rosenthal, 1996; Rosnow et al., 2000).

4.2.6.4 *Analysis of variance*

Two methods of reporting effect sizes for analyses of variance are Cohen's *f* (Cohen, 1988) and omega squared (ω^2). The former is the standard deviation of the group means relative to the common standard deviation of the groups, while the latter provides an unbiased estimate of the proportion of variance explained by the categorical variable. Cohen's *f* can be used for samples as well as for populations, while omega squared is a sample estimate of eta squared (η^2), the proportion of variance explained by population membership.

To calculate Cohen's *f*, a crude estimate of eta squared, $\hat{\eta}^2$ must be calculated as follows:

$$\hat{\eta}^2 = \frac{SS_{between}}{SS_{total}}$$

where $SS_{between}$ and SS_{total} are the between groups and total sum of squares.

Then, use the following formula to calculate Cohen's *f*:

$$f^2 = \frac{\hat{\eta}^2}{1-\hat{\eta}^2}$$

Calculate omega squared (ω^2) as follows:

$$\omega^2 = \frac{SS_{between} - (k-1)MS_{within}}{SS_{total} + MS_{within}},$$

where k = number of groups, MS_{within} = within groups mean square.

The sum of square and mean square information is provided by most statistical programs. After calculating Cohen's *f* or omega squared, use the descriptors in Table 3 to interpret these coefficients.

4.3 Confidence intervals

The reporting of confidence intervals around point estimates has been advocated as an alternative, or adjunct, to null hypothesis significance testing by many researchers, including both critics and proponents of null hypothesis significance testing (Cohen, 1990, 1994; Hunter, 1997; Schmidt, 1996; Steiger & Fouladi, 1997; Tukey, 1991). In this section an overview of confidence intervals is given, and the use of confidence intervals is evaluated.

4.3.1 An overview of confidence intervals

Confidence intervals provide different information from that arising from hypothesis tests. Hypothesis testing produces a decision about any observed difference: either that the difference is "statistically significant" or that it is "statistically non-significant" (Davies, 2001:3). In contrast, confidence intervals provide a range about the observed effect size. This range is constructed in such a way that we know how likely it is to capture the true – but unknown – effect size. Thus the formal definition of a confidence interval is: a range of values for a variable of interest constructed so that this range has a specified probability of including the true value of the variable. The specified probability is called the confidence level, and the end points of the confidence interval are called the confidence limits (Cumming & Finch, 2001:532).

It is a widespread convention to create confidence intervals at the 95% level; this means that 95% of the time properly constructed confidence intervals should contain the true value of the variable of interest. Hence, more colloquially, the confidence interval provides a range for the estimate of the true treatment effect that is plausible given the size of the difference actually observed (Davies, 2001).

By computing confidence intervals, it is possible to conclude that the real population value is somewhere between two values (the upper and lower confidence bounds), and the degree of confidence can also be specified (e.g., 80%, 90%, 99% confidence). For example, we could say that we are 90%

sure that tomorrow's temperature will be between 25 and 35 degrees, or we could say that we are 95% sure that the temperature will be between 20 and 40 degrees. In the second case, the confidence is higher, but the estimate is less precise. Determining which estimate is better is a matter of opinion (Cumming & Finch, 2001).

Another example is given to illustrate the point: If the researcher has a sample of 10 learners, and the mean time is 2.5 minutes for a reading task, computing a 90% confidence interval might lead to the following conclusion: I can be 90% sure that the mean time for the entire population of learners would be somewhere between 2 and 3 minutes. A 95% confidence interval might say the following: I can be 95% certain that the true population value is between 1.5 and 3.5 minutes. What the researcher really wants is to have high confidence that the population value falls somewhere in the interval estimate, and for the interval to be as precise (small) as possible. The way to get both of these is to increase the sample size, the larger the sample, the more precise the confidence intervals will be. For example, if the sample size is doubled, the 90% confidence interval might shrink to 2.3 and 2.7, and the 95% confidence interval might shrink to 1.9 and 3.1.

4.3.2 Calculating confidence intervals for a population

If the researcher knows the standard deviation of the population, it is possible to calculate the confidence intervals:

$$CI = \bar{x} \pm (Z_{\alpha/2})(\frac{\sigma}{\sqrt{n}})$$

This is the confidence interval for the population mean of x when the population standard deviation is known (σ) (Cumming & Finch, 2001; Lawrence, 2004). There are four important parts of the formula; these are discussed in the following sub-sections.

i) Standard error of the mean

One important part is $\sigma/\sqrt{n}$, which is called the standard error of the mean. The standard error of the mean is a measurement of how much error results from the size of the sample. The bigger the sample, the less likely researchers are to accidentally end up with an unrepresentative sample, and therefore, the standard error of the mean will be smaller. Conversely, if the sample is small, the chances of having a "bad" sample are higher, so the standard error will be bigger (Cumming & Finch, 2001). For example, if $\sigma = 1.3$ and $n = 144$, the standard error of the mean will be:

$\sigma/\sqrt{n} = 1.3/\sqrt{144} = 1.3/12 = 0.1083$

ii) Alpha and confidence levels

The next part of the formula is the confidence level. The confidence level represents how willing the researcher is to accidentally report a mistake. This is represented by $Z_{\alpha/2}$ in the equation. The Greek letter α actually represents something called the alpha level. Normally, the values $\alpha=0.05$ or $\alpha=0.01$ are used, which mean that the researcher is willing to be wrong five out of every hundred times, or one out of every hundred times, respectively (Hatch & Lazaraton, 1991; Cumming & Finch, 2001). However, instead of talking about α directly, the confidence level is stated in terms of the changes of NOT making a mistake. $\alpha = 0.05$ corresponds to a 95% confidence level, while $\alpha = 0.01$ corresponds to a 99% confidence level.

In the formula, two Z values (critical values) are of importance:

$Z_{0.05/2} = Z_{.025} = 1.96$ and

$Z_{0.01/2} = Z_{.005} = 2.58$

iii) Confidence limits

The confidence limits are the product of the standard error of the mean and the Z score for a given confidence level (Cumming & Finch, 2001).

$$e = (Z_{\alpha/2})(\frac{\sigma}{\sqrt{n}})$$

e represents the confidence limit for a given α, σ and n.

For the example given above, the confidence limits for the sample mean can be calculated, given n=144 and σ=1.3, as before, for α=0.05:

$$e = (Z_{\alpha/2})(\frac{\sigma}{\sqrt{n}}) = Z_{.025}\frac{1.3}{\sqrt{144}} = 1.96\times\frac{1.3}{12} = 1.96\times 0.10833 = 0.25483.$$

With $\alpha = .01$:

$$e = (Z_{\alpha/2})(\frac{\sigma}{\sqrt{n}}) = Z_{.005}\frac{1.3}{\sqrt{144}} = 2.58\times\frac{1.3}{12} = 2.58\times 0.10833 = 0.2795.$$

Note that the interval is bigger with the smaller alpha.

iv) Confidence interval

Once the confidence limits have been determined, the confidence interval can be calculated:

$$CI = \bar{x} \pm (Z_{\alpha/2})(\frac{\sigma}{\sqrt{n}})$$

If $\bar{x} = 0.20$, we can find the confidence intervals for $\alpha = .05$:

$$CI = \bar{x} \pm (Z_{\alpha/2})(\frac{\sigma}{\sqrt{n}}) = 0.20 \pm 0.2548 = (-0.0548 < \mu < 0.4548)$$

When $\alpha = .01$:

$$CI = \bar{x} \pm (Z_{\alpha/2})(\frac{\sigma}{\sqrt{n}}) = 0.20 \pm 0.2795 = (-0.0795 < \mu < 0.4795)$$

This means that the population mean is somewhere between -0.0548 and 0.4548, with a 95% confidence level, and between -0.0795 and 0.4795, with a 99% confidence level. To put it another way, a researcher belies that, with the given margin of error, the mean of the population is somewhere between those values, and that it would be possible to get a result between those values if the researcher considered the entire population 95% (99%) of the time.

4.3.3 Minimum sample sizes

From the confidence limit formula it is possible to derive the minimum n (sample size) required to have a given confidence level. That is, how many items out of the population must the researcher look at to be confident that the population parameter is close to the sample statistic. In order to do this, the confidence limit formula is rearranged (Lawrence, 2004):

$$n = (\frac{(Z_{\alpha/2})\sigma}{e})^2$$

For example, if the researcher wanted to know the average age of North-West University students within a month (1/12 of a year, or 0.0833 years), with a 95% confidence level (α=0.05), and the standard deviation is known σ=1.8:

$$n = (\frac{(Z_{\alpha/2})\sigma}{e})^2 = (\frac{1.96 \times 1.8}{0.0833})^2 = (\frac{3.528}{0.0833})^2 = (42.336)^2 = 1792$$

This means that the researcher would need to sample 1792 students to determine this information with a 95% confidence level.

4.3.4 Calculating confidence intervals: Population standard deviation unknown

Most of the time, researchers do not already know the population standard deviation (σ). However, there is an analogous formula that researchers can use with the sample standard deviation, s (Lawrence, 2004):

$$CI = \bar{x} \pm (t_{\alpha/2})(\frac{s}{\sqrt{n}})$$

In this formula σ has changed to s, and Z has changed into t.

The new distribution is called the student's t distribution. The student's t distribution is much like the standard normal curve, except that it is designed specifically to deal with the problem of small samples. The degrees of freedom are df=n-1 for the t distribution. For any given α and df (or n), the researcher can find the correct t value. For example, the t value for α=0.05 and 18 degrees of freedom will be t_{crit}=2.101.

Once the t distribution is know, the researcher can calculate the confidence intervals. For example, let us assume the researcher has a sample of 20 undergraduates majoring in English at North-West University. Their mean IQ x=107 with a standard deviation s=17. What is the confidence interval for all undergraduate English majors at North-West University, with a 95% confidence level?

Since n=20, n-1=19 degrees of freedom. The critical value for t with α=0.05 and df=19 is 2.093.

$$
\begin{aligned}
CI &= \bar{x} \pm (t_{\alpha/2})(\frac{s}{\sqrt{n}}) = 107 \pm (2.093)(\frac{17}{\sqrt{20}}) \\
&= 107 \pm (2.093)\frac{17}{4.47} \\
&= 107 \pm (2.093)(3.081) = 107 \pm 7.956 \\
&= 99.04 < \mu < 114.96
\end{aligned}
$$

4.3.5 Evaluation of confidence intervals

Loftus (1991) has also been a strong advocate of using confidence intervals in reporting results. For him, the provision of intervals constitutes a useful and necessary means of accounting for error and provides the researcher with a much more meaningful answer to the question one is *really* probing:

> The emphasis on hypothesis testing produces a concomitant de-emphasis of an alternative technique for coping with statistical error that is simple, direct, and intuitive and that has wide acceptance in the natural sciences: the use of confidence intervals. Whereas hypothesis testing emphasizes a very narrow question ("Do the population means fail to conform to a specific pattern?), the use of confidence intervals emphasizes a much broader question ("What *are* the population means?"). Knowing what the means are, of course, implies knowing whether they fail to conform to a specific pattern, although the reverse is not true. In this sense, use of confidence intervals subsumes the process of hypothesis testing (Loftus, 1991: 103).

Confidence intervals allow the researcher to estimate the degree to which (i.e., probability) the observed value (or difference) is likely to be the *true* value. What makes confidence intervals so appealing is that they can provide both a hypothesis test, and produce an estimate of the observed parameter

(Becker, 1991). This is true despite the fact that the hypothesis test and confidence interval are still regarded as separate steps in the process of data appraisal (Serlin, 1993).

One useful feature of confidence intervals is that one can easily tell whether or not statistical significance has been reached, just as in a hypothesis test.

- If the confidence interval captures the value reflecting "no effect", this represents a difference that is statistically non-significant (for a 95% confidence interval, this is non-significance at the 5% level);
- If the confidence interval does not enclose the value reflecting "no effect", this represents a difference that is statistically significant (for a 95% confidence interval, this is significance at the 5% level) (Davies, 2001:4).

Thus, statistical significance can be inferred from confidence intervals, but, in addition, these intervals show the largest and smallest effects that are likely given the observed data. This is useful extra information.

One of the advantages of confidence intervals over traditional hypothesis testing is the additional information that they convey. The upper and lower bounds of the interval give us information on how big or small the true effect might plausibly be, and the width of the confidence interval also conveys some useful information.

If the confidence interval is narrow, capturing only a small range of effect sizes, we can be quite confident that any effects far from this range have been ruled out by the study. This situation usually arises when the size of the study is quite large and hence the estimate of the true effect is quite precise. However, if the confidence interval is quite wide, capturing a diverse range of

effect sizes, we can infer that the study was probably quite small. Thus any estimates of effect size will be quite imprecise (Davies, 2001).

Some researchers have cautioned that although confidence intervals can be very effective aids to the interpretation of data, they are subject to some of the same types of misconceptions and misuses as is null hypothesis testing (Cortina & Dunlap, 1997; Frick, 1996; Hayes, 1998). Steiger and Fouladi (1997) warned that the proper use of confidence intervals requires assumptions about distributions, just as null hypothesis significance testing does, and that if the assumptions are not met the resulting intervals can be inaccurate. In particular, they noted that "the width of a confidence interval is generally a random variable, subject to sampling fluctuations of its own, and may be too unreliable at small sample sizes to be useful for some purposes" (Steiger & Fouladi, 1997: 254).

4.4 Power analysis

The purpose of many statistical tests is to decide, for a given research hypothesis, whether the researcher should reject the null hypothesis that the treatment being tested has no effect. When making this decision, it is possible to make two types of decision errors. A Type I error involves rejecting the null hypothesis when it is actually true, i.e., concluding that there is a statistically significant difference between the treatment groups in an experiment when no true difference exists. The probability of a Type I error, termed *alpha*, is set by the researcher for a given study (e.g., $P = 0.05$). The second type of decision error, a Type II error, occurs when one accepts the null hypothesis when it is actually false, i.e., when one concludes that there is no difference between the groups tested in an experiment, when a true difference exists. The probability of a Type II error is termed *beta*. The probabilities of Type I and Type II errors are inversely related; as alpha increases, beta decreases, and vice versa (Hatch & Lazaraton, 1991).

Statistical power is the ability to detect a difference between treatment groups in an experiment, if a true difference in means exists, i.e. the probability to reject the null hypothesis if a given difference in means exists. When the statistical power of an experiment is high enough, it is possible to obtain a statistically significant result if there is a true difference. If power is not high enough, a study may not produce a statistically significant result even when there is a true difference. Power is largely determined by two factors (Aron & Aron, 2003). The first, effect size, is the strength of the association within the population between the effect of interest and the dependent variable. The larger the effect size, the greater the power, and vice versa. Effect size is influenced by the standard deviation of the dependent variable within the population. The larger the standard deviation, the smaller the effect size and the lower the power; and vice versa, the lower the standard deviation, the larger the effect size and the higher the power. The second important factor in determining power is the number of participants in the study. When other factors are held constant, increasing the number of participants increases the experiment's power, and reducing the number of participants reduces the power. Power is also influenced by the alpha level selected by the researchers as well as the kind of statistical test used in the study (Aron & Aron, 2003). Power can vary from 0 to 1 (0 to 100 percent). Cohen (1988; 1992) recommends that experiments be designed to achieve a power of about 0.80 (80 percent). When alpha is 0.05, a power of 0.80 is associated with a 20 percent probability of a Type II error and 5 percent probability of a Type I error.

In power analysis, one uses the relationships between power, effect size, and number of participants in designing experiments and interpreting results. For example, by knowing (or predicting) the effect size, and choosing the desired power and alpha level, it is possible to estimate the number of participants needed in a study. For studies for which effect sizes are expected to be larger, the use of power analysis would help researchers to avoid including more participants than needed. Because increasing the number of participants in an experiment increases the chances of obtaining statistically significant results, even a tiny effect of little practical importance can be found

to be statistically significant if an experiment includes too many participants (Murray & Dosser, 1987). One way to roughly assess the practical significance of an effect is to compare an estimate of the effect size to known conventions. Effect size can be estimated in a variety of ways, depending on the type of statistical analysis used in an experiment (Cohen, 1992 presents a table of effect size measures for eight common statistical tests). When, for example, one-way analyses of variance (ANOVAs) are used, the usual measure of effect size is eta-squared (η^2), defined as the proportion of the variance on the dependent measure accounted for by the independent variable (eta-squared is computed by dividing the sum of squares for the effect of interest by the total sum of squares). Theoretically, eta-squared can vary from 0 to 1, but in practice, it rarely reaches 0.5 (Aron & Aron, 2003). Values for eta-squared approximately correspond to the following effect-size conventions: small (0.01), medium (0.06), and large (0.14) (Cohen, 1988).

Completing a power analysis and estimating effect sizes for dependent variables can prove to be useful in several respects. First, it becomes possible to judge whether statistically non-significant results are more likely due to a study's small sample size than to the absence of true effects. It is possible to make this judgment by (a), estimating effect sizes, and then (b), estimating the samples sizes required to obtain adequate power, given these effect sizes. Second, power analysis makes it possible to judge whether redesigning and rerunning a study could increase its power and the likelihood of obtaining significant results if true between-group differences exist. It is possible to make this judgment using (a) estimates of required sample sizes, to estimate the number of additional participants needed to attain adequate power for each dependent variable, and (b) knowledge or informed guesswork about the degree to which redesign measures could boost participation.

However, like all statistical tools in the researchers' kit, power analysis has its limitations. How likely an experiment is to yield statistically significant results depends not only on the size of the effect (assuming one exists) and the size of the sample but also on the within- condition variability of the data. Within-condition variability can be influenced by many factors, but the more it can be

limited by careful experimental control of extraneous variables, the more likely an effect of a given size is to show up as statistically significant, other things being equal; and power analysis is not sensitive to this fact. It should be borne in mind, however, that attempts to minimize within-group variability can have the effect of limiting the generality of findings (Pruzek, 1997); findings obtained with relatively homogeneous samples (e.g., university first-year students) do not necessarily generalize readily to more heterogeneous populations (e.g., the general public).

4.5 Conclusion

Each of the suggested alternatives or complements to null hypothesis significance testing mentioned above has its proponents. In trying to assess the merits of null hypothesis significance testing and other approaches to the evaluation of hypotheses, it is well to bear in mind that nothing of importance in reading instruction research has ever been decided on the basis of the outcome of a single statistical significance test. Reading instruction knowledge is acquired as a consequence of the cumulative effect of many experiments and non-experimental observations as well. It is the preponderance of evidence gathered from many sources and over an extended period of time that determines the degree of credibility that is given to hypotheses, models and theories.

There are no statistical procedures that can safely be used without thought as to their appropriateness to the situation or the question of whether there are more effective approaches available.

CHAPTER 5
METHOD OF RESEARCH

5.1 Introduction

The aim of this chapter is to discuss the methodology employed in this study. The methodology consists of a detailed review of the literature focusing specifically on ESL/EFL reading instruction and the statistical reporting of the data. The methods for this literature review are based on four interconnected steps: (i) the choice of content-based and methodological inclusion and exclusion criteria for the studies to be used in the review; (ii) the identification of these studies; (iii) the extraction of information from these studies; and (iv) the methods for synthesising this information.

5.2 Inclusion and exclusion criteria

The review and integration of research literature begins with the identification of the literature. Locating studies is the stage at which the most serious form of bias enters a study (Glass, McGraw, & Smith, 1981). Glass (1976:6) states that: "How one searches determines what one finds; and what one finds is the basis of the conclusions of one's integration". The best protection against this source of bias is a thorough description of the procedure used to locate the studies.

Studies were included using content-based and methodological criteria. For a study to be included, it had to be a study looking at the effect of reading instruction/reading instruction programme (e.g., phonemic awareness, strategy teaching methods, etc.) on learners' reading achievement (e.g., vocabulary acquisition, fluency or reading comprehension). Control treatments could take any form, such as routine classroom teaching of reading, placebo computer teaching, or other non-computer teaching of reading. Studies focusing on the relationship between reading strategy use by ESL/EFL learners and learning

styles or reading achievement, etc. were therefore excluded.

As the focus of the review is "evidence-based research", studies using rigorous methods to assess effectiveness were required. In essence, this implied randomised controlled trials (RCTs). These use an 'experimental' design in which the participants are divided randomly into experimental and control groups, with the experimental group receiving the intervention (reading instruction of some kind), whilst the control group does not receive this intervention. The simple elegance of this design is compromised if the participants are not *randomly assigned* to the two groups. If participants are allocated on any other basis, one cannot be sure whether (except for chance differences) the experimental and control groups were similar before receiving (or not receiving) the intervention. Whereas one can use statistical techniques in an attempt to control for the potential confounding from variables which one knows to be prognostic in terms of their effects on the outcomes being considered, these techniques are far from perfect and, crucially, cannot adjust for unknown variables. Thus, non-randomised designs cannot be certain to generate groups which do not differ (except by chance) in either known or unknown factors, and hence these designs always have the potential that selection biases may make comparisons between the two groups about the effect of the intervention uncertain.

For this review, because of the relatively few studies in reading instruction research utilizing RCTs, studies focusing on RCTs as well as studies employing quasi-experimental designs with control groups, were analysed and evaluated.

5.3 Identification of studies

This process had two steps: (a) the search strategy for identifying likely studies; and b) the strategies for deciding to include the identified studies.

Between 16 and 22 March 2002 and then again between 1 and 30 November

2005, five bibliographic databases were electronically searched for RCTs: The Educational Research Information Clearinghouse (ERIC); The British Educational Index (BEI); PsycInfo; Web of Science (Science and Social Science); and EBSCOhost. The searches were for the period 2000-2005. As the focus was ESL/EFL reading instruction, the search was restricted to the English language literature. It is not easy to identify RCTs from educational databases. Indeed, even in databases of psychological literature, RCTs are not systematically key-worded. To address this difficulty, a very sensitive, but not very specific, search strategy was used. This maximised the chance of identifying all relevant studies, but at the cost of identifying many irrelevant studies. The eventual search strategy used the key words of reading and instruction, combined with the words comprehension* random* and experiment*.

In addition to the bibliographic databases that were searched six journals that were readily accessible online to practising English educators were randomly selected for inclusion purposes. These journals included: *The Reading Matrix, Language Learning and Technology, Scientific Studies of Reading, System, Teaching English as a Second or Foreign Language* and the *Journal for Language Teaching* (South Africa). There was a great deal of overlap between the studies selected from the databases and the online journals.

The titles and abstracts of a total of 583 studies were reviewed for the purpose of inclusion and analysis in this study (cf. Table 4).

Table 4: Number of titles and abstracts reviewed

Journals	Total
The Reading Matrix	76
Language Learning and Technology	89
Scientific Studies of Reading	98
System	180
TESL-EJ	59
Journal for Language Teaching	81
TOTAL	**583**

A total number of 16 studies met the inclusion criteria (cf. section 5.2) and were selected for final analysis purposes (cf. Chapter 6). The studies were initially screened by the researcher using titles and abstracts only. A 10% random sample was also screened by the researcher's promoter. All screening was undertaken independently. Any discrepancies *were* then discussed and resolved. The full papers were retrieved where there was agreement on their inclusion in the in-depth review. In the event of continuing disagreement at the screening stage, the full paper was retrieved anyway. To check for reliability of screening procedures, a score for inter-rater reliability was calculated. A cut-off date of 5 December 2005 for receipt of the papers was selected on pragmatic grounds.

5.4 Data extraction

Using other literature reviews as a guide (Albanese & Mitchell, 1993; Dochy, Segers & Buehl, 1999; Vernon & Blake, 1993), the researcher defined the characteristics central to the review and analyzed the articles that were selected on the basis of these characteristics. Specifically, the following information was recorded in tables (cf. Chapter 6):

- Author(s) and year of publication;
- Journal in which article was published;
- Research questions and/or hypotheses;

- Subjects/Participants - countries, number of different schools, number of different classrooms, number of participants, age, grade, reading levels of participants, location (urban, rural, suburban), SES, ethnicity, exceptional learning characteristics (reading disabled, etc.), and selection criteria;
- Design - random assignment of participants to treatments with or without a pretest, treatment components administered in a fixed vs. variable order, baseline description;
- Description of treatment and control conditions - nature and components of reading instruction provided to control group, implicit or explicit instruction, difficulty level of any text used, duration of treatment (minutes per session, sessions per week, and number of weeks);
- Data analysis techniques and statistics reported – how was the data analysed? What techniques were used and what statistics were reported?
- Results of the study.

Data about the learners in the trials, about the outcomes of the interventions, and about the quality of the studies, were extracted by the researcher from all the included papers, using a standard format. Data from 50% of these papers were also extracted in the same way by the promoter. Any discrepancies were then discussed and resolved.

5.5 Synthesis

Two commonly used methods to review literature are narrative reviews and quantitative methods. In a narrative review (the focus of this study), the researcher tries to make sense of the literature in a systematic and creative way (Van Ijzendoorn, 1997). Quantitative methods utilize elementary statistical procedures for synthesizing research studies (e.g., counting frequencies into

box scores). These methods are more objective but give less in-depth information than a narrative review (Dochy, Segers, & Buehl, 1999).

The researcher explored the methodological quality of RCTs, using the checklist in the CONSORT guidelines for reporting RCTs in health care (Moher *et al.*, 2001). The researcher focused particularly on the internal validity of the studies. To limit the possibility of introducing reviewer bias, the studies were assessed for quality before the outcome data were extracted.

5.6 Conclusion

The essence of the methodology employed in this study was an analysis of the literature focusing on ESL/EFL reading instruction research and specifically the statistical reporting practices utilised in these studies.

CHAPTER 6

PRESENTATION AND DISCUSSION OF RESULTS

6.1 Introduction

The purpose of this chapter is to address the following research question as formulated in chapter 1:

- What is the state of affairs with regard to statistical significance testing in reading instruction research, with specific reference to post-1999 literature?

6.2 Results

Of the journals reviewed in this study, the following journals, namely *Language Learning and Technology*, *Scientific Studies of Reading* and *Teaching English as a Second or Foreign Language* utilise the Publication Manual of the American Psychological Association (1994; 1995) as a source of clear communication for authors submitting manuscripts. The American Psychological Association (1994:16) "encouraged" that authors of manuscripts using inferential statistics "include sufficient information to help the reader corroborate the analyses conducted". Moreover, when reporting inferential statistics, authors of manuscripts should "include information about the obtained magnitude or value of the test" (The American Psychological Association (1994:15).

Maxwell, Camp, and Arvey (1981:525) suggested that the "primary advantage of measures of strength of association is that they have the potential to reveal whether a statistically significant result reflects a meaningful rather than a trivial experimental effect". Critics see statistical significance testing as nothing more than a numbers game where researchers are only concerned with reporting statistically significant results even when the results were not of any practical importance (Daniel, 1997; Thompson, 1987; Vacha-Haase & Nilsson, 1998).

In order to determine if the statistical significance tests are of any practical significance, Vasquez, Gangstead, and Henson (2000) reiterated that journals in the education field require authors to report the relative treatment magnitude along with the statistical significance test.

While researchers solely rely on statistical significance testing, either using the observed significance level (P-value) or test statistics like F, t, or x^2, the researcher may be distracted from more important considerations like result importance or value, result replication, and result magnitude or effect (McLean & Ernest, 1998; Thompson, 1999a, 1999b, 1999c).

6.2.1 Journal analyses

In this section the studies selected for analysis purposes are discussed. The studies are discussed under the journal headings in which they were published. None of the studies published in the *Journal for Language Teaching* met the inclusion criteria as specified in chapter 5.

6.2.1.1 Studies in The Reading Matrix

In the study conducted by Bell (2001) (cf. Table 5), the design was not specifically stated, but one is left to deduce that it is a pre-test post-test control group design. There is, however, no indication whether the participants were randomly assigned to the treatment conditions. The researcher used t-tests to analyse the data even though the two groups were very small (n=14 and n=12). No mention is made of the normality of the data distribution and it would have been more appropriate to use non-parametric statistics to analyse the data. No effect sizes are reported to determine whether the results are of any practical significance. The two hypotheses formulated for the study were accepted: "The hypothesis that learners in the extensive group would achieve **significantly faster** reading speeds than subjects in the intensive group is very **strongly supported** by the data" (my emphasis) (Bell, 2001:9); "The results of the comprehension tests also provide **very strong support** for the hypothesis that learners in the extensive group would achieve **significantly higher** scores than learners in the intensive group" (my emphasis) (Bell, 2001:10). The typical use of words such

as significantly instead of statistically significantly is used throughout the discussion of the results. Thompson and Snyder (1997) state that researchers use language like "significance" when they meant "statistically significant" resulting in misleading uses of the wording.

Limitations of this study include the small number of subjects as well as possibly the Hawthorne Effect, Hallo Effect, and Subject Expectancy (cf. Brown, 1988: 33-34). The researcher also states that the subjects in the extensive group were aware of being involved in a separate and special reading programme, and the high profile of the research could have led subjects to "assist" the researcher by modifying the motivation of learners in the extensive group, and therefore their performance on the tests (Bell, 2001:10).

The results of the study conducted by Stakhnevich (2002) indicate that the medium of instruction does have an impact on the level of reading comprehension, with the web mode resulting in better performance when compared to the traditional print mode (cf. Table 5). Based on the statistical significance of the results, the researcher rejects the null hypothesis formulated for the study. Limitations of the study include the relatively small sample size (n=11, n=10 and n=10, respectively for the web group, the traditional print group, and the control group) and the use of parametric statistics when it might have been more useful to utilise more conservative non-parametric tests. The researcher makes the following conclusion based on the results: "For ESL educators these results suggest that when teaching content through self-directed reading, the web medium can evoke better reading comprehension than the traditional print medium" (Stakhnevich, 2002:13).

In the study conducted by Johnson and Howard (2003), they used a one-group pre-test post-test design with intact groups of classes (cf. Table 5). Randomisation was not done and a control group was not used. According to Hatch and Lazaraton (1991), this is one of the weakest designs researchers can use when conducting research studies. Nevertheless, Johnson and

Howard (2003:90) make the following conclusion based on the results: "The AR program was effective in improving the reading skills of urban, inner-city students". The use of statistical methods would only have been appropriate if the group was a random sample. Usually such large samples result in statistically significant results which can have low effect size.

In his study, Gupta (2004) did not specify the research design followed in his study (cf. Table 5). He did, however, specify that the subjects were randomly assigned to the experimental and control groups. The researcher used percentage gains on test scores to compare the performance between the experimental and control groups after intervention. Data were also qualitatively reported by means of narratives. No parametric statistics were used and the researcher did not conduct effect size calculations to determine whether the intervention was of any practical significance. The researcher does, however, make the following conclusion: "The results of this research **clearly indicate** that clinical intervention makes a difference for the development and outcomes of reading skills of children" (my emphasis) (Gupta, 2004:61). The researcher makes the following comments with regard to the qualitative data: "However, it is the qualitative affective aspects of intervention results in challenging settings that are of more human value. Specifically, the improvement in children's attitude towards reading, willingness to read on their own, gain in their confidence and trust is observed by their teachers and parents ..." (Gupta, 2004:60).

In all of the above-mentioned studies the methodologies seem to be flawed in one way or another. The researchers all rely very heavily on the statistical significance of their results and based on this they reject the formulated null hypotheses. In none of the studies reviewed in *The Reading Matrix* did the authors make use of effect size measures to indicate the practical significance of their results.

Table 5: Studies in The Reading Matrix

Authors	Bell (2001)
Research questions and/or hypotheses	H_1: Learners in the extensive group will achieve significantly faster reading speeds than those in the intensive group as measured on relatively easy, non-problematic texts. H_2: Learners in the extensive group will achieve significantly higher scores on a test of reading comprehension containing texts at an appropriate level, than those in the intensive group.
Participants	Two groups of elementary level learners (n=26) at the British Council English Language Centre in Sana'a, Yemen.
Design	Pre-test post-test control group design
Data analysis techniques and statistics reported	T-test for correlated samples (to compare scores prior to and after the study on the same group); t-test for independent samples (to compare performance between groups); t-value (t=7.14*** df=11); statistical significance (*** p<0.001); df; mean gain scores and percentages
Intervention	The experimental group (n=14) received an extensive reading programme consisting of class readers, a class library of books for students to borrow, and regular visits to the library proving access to a much larger collection of graded readers (up to 2000 titles). The study extended over a period of two semesters and the reading programme covered one quarter of the total class time (36 out of 144 hours). An inventory of readers was compiled, and reading records maintained to record titles read and the time subjects spent reading each week.
Control	The control group (n=12) received an entirely different reading programme which was intensive in character, being based on the reading of short passages and the completion of tasks designed to "milk" the texts for grammar, lexis, and rhetorical patterns. For this purpose, the control group made a detailed study of the title "Basic Comprehension Passages" by Donn Byrne. This text contains thirty short texts of around 300 words each, arranged in groups of ten, together with a wide variety of exercise types for intensive exploitation of the passages.

Results	Subjects exposed to extensive reading achieved both significantly faster reading speeds and significantly higher scores on measures of reading comprehension.
Authors	Stakhnevich (2002)
Research questions and/or hypotheses	H_1: There will not be a significant difference between the posttest scores of the ESL students from the two treatment groups and the posttest scores of the ESL students from the Control group; H_2: There will not be a significant difference between the posttest scores of the ESL students from the Web group and the posttest scores of the ESL students from the Traditional print group; H_3: There will not be a significant difference between the posttest scores of the ESL students with a higher TOEFL score and the posttest scores of the ESL students with a lower TOEFL score.
Participants	The population for this study was defined as all newly enrolled ESL students who came to the United States one to two weeks prior to the beginning of this research project and began attending an American medium-size public university or its Intensive English Programme during the fall semester of 1999. The population consisted of ninety students. The size of the sample was 31 voluntary participants, including 12 Chinese students, 5 students from the former Soviet Union, 4 students from Malaysia, 2 from Germany, and one from each of the following countries: Nepal, India, Saudi Arabia, Thailand, Japan, and Venezuela. All participants were over 18 years old and for all of them this was their first visit to the United States. None of the participants had previously lived in an English-speaking country. The gender distribution of the sample was 16 female and 14 male subjects. The subjects' TOEFL score ranged from 400 to 647 with 597 being the median. The subjects were divided into two subgroups: those with a higher TOEFL score (scores higher than and equal to 597) and those with a lower TOEFL score (scores below 597). Then the subjects from each subgroup were randomly assigned to the Web group, Traditional print group or Control group.
Design	2X3 factorial design. The participants were randomly assigned to the Control, Traditional print and Web groups.
Data analysis techniques	Descriptive statistics (e.g., Means and standard deviations) ANCOVA was employed to test H_1 and H_2.

and statistics reported	Factorial ANOVA was utilised to test H_3. F-values and *df* are reported as well as statistical significance ($P<0.05$)
Intervention	Students who were placed in the **Web group** were asked to remain in the lab and work with the Mississippi Web ESL Project using the web medium. They were shown how to navigate the web, use online glosses and the online version of the Merriam-Webster dictionary. They were encouraged to practise with the computer before starting to read and ask questions if they had technical problems while on task. It took approximately one hour and twenty minutes for everyone in the Web group to complete the assignment. The **Traditional print group** participants were taken to a nearby classroom and presented with a book version of the Merriam-Webster dictionary and a print version of the Mississippi Web ESL Project. They completed their reading in approximately an hour and ten minutes.
Control	Students who were placed into the **Control group** were taken to a nearby classroom and asked to watch a segment from "Ferris Bueller's Day Off" for approximately an hour. Afterwards, the students were encouraged to participate in a teacher-led discussion of the differences and similarities between high school education in the United States and in their native countries. The discussion lasted approximately half an hour and as indicated by several participants left them with a better understanding of American high school education. The video segment and the topics that emerged in the classroom discussion were unrelated to the topics covered in the Mississippi Web ESL Project and in this way did not have any inherent learning value pertinent to this research.
Results	The null hypotheses were rejected. The results of the study suggest that when teaching content through self-directed reading, a web medium can evoke better reading comprehension than the traditional print medium. The ESL students with a higher TOEFL score did not show a significantly higher level of reading comprehension of the text when compared to those ESL students with a lower TOEFL score.
Authors	Johnson & Howard (2003)
Research questions	The purpose of this study was to determine whether the AR programme was effective, and whether its effectiveness was dependent upon the amount of student usage. More specifically, did students classified as "Low" users and "Average" users gain

and/or hypotheses	as much on reading comprehension measures as students classified as "High" users?
Participants	All of the subjects (755) in the current study attended seven inner-city, Title 1, elementary schools. One hundred and sixty-six third graders, 297 fourth graders, and 292 fifth graders participated in the year long AR programme. The majority of the students in the selected schools were considered at-risk as they qualified for the free lunch programme. Eighty-five percent of the students were African-American and approximately fifty percent were males.
Design	This study used a one-group pre-test post-test design with intact groups of classes. Randomization of subjects was not possible and a control group was not available.
Data analysis techniques and statistics reported	Student reading comprehension and vocabulary was examined using a Multivariate Analysis of Variance (MANOVA). The Wilks'-Lambda test was used as a test of significance for the MANOVA.
Intervention	Students that participate in the AR programme choose their own books from the Accelerated Reader book list, which contains more than 12,000 titles. Each of the 12,000 books is assigned a point value based on its length and the Flesch-Kincaid reading index to determine readability (Flesch, 1974:23). Students read selected books at their own pace and then take a test on the computer. The computer test consists of multiple choice questions about important facts in the book. Most of the questions evaluate literal comprehension. In order to earn any points on a book, the student must answer at least 60 percent of the questions correctly on the test. Careful test writing and security features in the software greatly reduce the possibility of student cheating. AR points are therefore a fairly accurate measure of the quantity of words being read and comprehended. Students may only test once for a given book. If students read too quickly, they score poorly because they are not reading with comprehension. When implemented according to design, teachers are expected to oversee students' reading patterns. If a child's test scores are too low, teachers intervene with advice on reading level and rates. The computer scores the test, calculates the number of points earned by each student, and records the data. Reports are generated listing the AR points

	earned, number of tests taken, number passed, average grade level of books read, and average percentage achieved on the tests taken. The treatment consists of three types of AR usage: Low participation (0- 20 AR points); Average participation (21- 74 points); and High participation (75 points and above). Generally, students with students with Low participation read less than three books during an academic year. Average participants read three to five books, while High participants read more than eight to ten books.
Control	No control group was used.
Results	The AR programme was effective in improving the reading skills of urban, inner-city students. Participants in all three usage groups improved their reading skills as measured by the *Gates-MacGinitie Reading Test.* In statistical language significant main effects occurred in concurrence with the degree of AR usage (Wilks' Lambda = .9428, F = 11.13, $p < .0001$). Students who read the most (High Participants) gained 2.24 years on the *Gates-MacGinitie Reading Test.* The Average Participants gained 1.52 years; and the Low Participants gained .73 of a year. The High AR Participants exceeded normal expectations for gains by an additional one year and two months. The Average Participants gained an extra one half year, while the Low Participants only achieved three quarters of an academic year's progress. Thus, as reading practice increased with AR usage, reading comprehension and reading vocabulary scores improved. Overall, there were low levels of AR participation among the 755 students in the current study. Fifty-two percent of the students earned less than 21 points, the equivalent of reading two or three books. They were classified as "Low AR Users." Thirty-six percent of the sample, labeled as "Average Usage", earned between 21 and 74 points. Less than 12 percent of the sample, the "High Usage" participants, earned more than 75 points. This indicates that most of the students in this investigation were reading much less than the recommended one hour per day.
Authors	Gupta (2004)
Research	Did on-site clinical intervention make a difference in the reading scores of tutored children compared with non-tutored children?

questions and/or hypotheses	Children who received tutoring from graduate Reading Interns (experimental group) would show greater growth over one semester of individualised tutoring relative to children who did not receive tutoring services (control group).
Participants	The inner-city urban elementary school participating in the programme had economically disadvantaged children with low achievement – majority of the children (over 95%) were on free or reduced lunch programmes. The school was in the bottom quartile in the district based on overall academic scores ranking. Participants were children in third and fifth grades who were at the bottom of the quartile in each grade based on test scores. The children selected to participate in the study were selected by using criteria such as: teacher input, letter grade in reading, test scores, and teacher observation of their reading ability. For the study, 40 children were selected from grade 3 and 5 across different sections. Since there were 17 graduate reading interns for one-one tutoring, 17 (randomly selected) of the 40 children were placed in the "experimental group", who received tutoring. The other 23 children formed the control group.
Design	Children randomly divided into two groups.
Data analysis techniques and statistics reported	The two groups were compared using percentages Qualitative data reporting based on surveys
Intervention	The children met one-on-one after school with reading interns once a week for a total of 13 sessions from January to April. The clinical intervention/tutoring took place on-site (in the school, as opposed to traditional reading clinics at a university campus) in a one-on-one supervised setting. Each session lasted for an hour. Interns followed clinical lesson plans during the session. Initial sessions focused on learning about tutees' strengths and needs. Because there is much variance in the reading strategies of children and the instructional paths leading to improved reading, the information needed in reading diagnosis varies from child to child and clinician to clinician. Depending on the individual learner,

	diagnosis typically included (but not limited to) the following: graded word recognition, reading inventory (includes, word recognition, miscue analysis, oral and silent reading evaluation, comprehension assessments for narrative and expository texts, reading rate, listening comprehension, summary of student's performance), administration of a standardized test, such as, Woodcock Reading Mastery Test or Peabody Picture Vocabulary Test, Phoneme Awareness (Spelling) (Johns, 2001), Phoneme Segmentation (Johns, 2001), Writing (Johns, 2001). A sample of suggested instructional activities included the following elements: Word Activities (word sort, concept sort, sound sort, open sort, sentence building), Pre-Reading Activities (previewing, predicting, picture walk, activating prior knowledge), During Reading Activities (guided reading, monitoring reading, shared reading, echo reading, alternate reading, assisted reading, independent reading), Post Reading Activities (retelling, main ideas, summarizing, connecting with student's life), Identifying Words (alphabetic principle, phonemic awareness, segmenting and blending phonemes, graphophonics, initial and final sounds, consonant clusters, vowels, syllables, onset, rime patterns, structural analysis, word patterns, sight vocabulary, using context to predict words, CSS), Writing (drafts, topic journals, LEA, editing, spelling patterns, response to literature), Comprehension (retelling, questioning, literal, inferential, understanding fiction and non-fictional texts, , making connections, processing texts, K-W-L, DRTA, establishing purpose), Fluency (phrasing, repeated reading), Developing vocabulary (word map, semantic maps, graphic-organizer, using context, word parts, classifying, grouping). Assessment must inform instruction. Instructional intervention typically included but was not limited to: alphabetic principle, phonemic awareness, concept of word, guided reading at instructional level; word study; developing vocabulary, promoting comprehension, organizing information, study skills, fluency training; reading to/with the child and interactive writing. Intervention was specific to individual assessment results and needs of the client. Together, the supervisor and clinician ensured that cultural/linguistic/physical/emotional/and social needs of the learners were addressed in clinical sessions. Interns were required to meet the specific clinical sessions recommended in the course. They were observed and provided written evaluation of the observed session by the supervisors. Reading interns diagnosed each tutee and developed a specific plan of remedial instruction to address individual needs. Ongoing assessment guided the individualized instruction tailored to each child's specific strengths and needs. Classroom teacher input

	and parent input was constantly sought by the reading interns to further assist tutees. Assessment, evaluation and observational data guided instruction in subsequent sessions. Instruction was adjusted as per the results of the on-going assessments.
Control	No tutoring.
Results	The children who received one-on-one, after school clinical reading sessions outperformed the control group and their own previous performance in terms of both test letter grades and test scores. Overall, 76% of tutored children improved their letter grade in reading from the beginning of the school year to the end of the year grade in reading, compared with 35% of children in the non-tutored group. None of the children in the tutored group dropped a letter grade, while 17.64% of children in the non-tutored group did experience a decrease in their letter grade. Similarly, 56% of the children in the tutored group showed improvement in their STAR scores from fall 2001 to spring 2002, as compared with approximately 50% of the children in the non-tutored group.

6.2.1.2 *Studies in Language Learning and Technology*

In his study Al-Seghayer (2001) formulated a null hypothesis of no difference between video clip annotations and still picture annotations (cf. Table 6). At the outset of the study the author already states that: "it was anticipated that the participants' performance on the vocabulary test will be significantly higher for words that had been annotated with video clips than for words annotated with pictures" (Al-Seghayer, 2001: 211). For this study, a convenience sample was selected and the author utilised a within-subject design; each participant served as his/her own control. Steyn (2005) is of the opinion that this convenience sample is really a population since it is not apparent to what population the results have to be generalised. The author also uses the non-parametric technique, namely the Friedman test, to analyse the data. He justifies his choice of using a non-parametric statistical technique: "First, the distribution of scores across the annotation modes did not meet the normality assumptions required for ANOVA. Second, the number of score levels obtained within each annotation mode was not sufficiently continuous for an analysis-of-variance approach" (Al-Seghayer, 2001:219-220). The results of the study indicate that the null hypothesis is rejected based on the **"significantly better"** (my emphasis) performance of the participants on the text-plus-video condition than on their performance on the text-plus-picture condition. The author does not calculate the effect size of the results. The assumption that the convenience sample can be viewed as a random one from some specified population ought to be made, otherwise it is not clear how the significance obtained, can be generalised. The author states that the results of the study have the following pedagogical implications: "What has been presented above demonstrates that exposing learners to multiple modalities of presentation (i.e., printed text, sound, picture, or video) produces a language learning environment which can have a real impact on learning" (Al-Seghayer, 2001:226). In this study Al-Seghayer combines quantitative and qualitative methods of data analysis supporting Pressley's (2003) plea for not focussing exclusively on experimental research and ignoring qualitative research.

Nikolova (2002:114) states that: "The main question of the study, whether students learn vocabulary better if they participate in multimedia authoring, was not answered in an unequivocal way. On the one hand, if time on task was taken into account the answer was negative. On the other hand, if time on task was disregarded, the authoring treatment yielded significantly better results in vocabulary acquisition". She continues by stating that, "the results of the study should be interpreted cautiously because of the small numerical difference in the scores of the experimental and the control groups, the small sample size, and the fact that the authoring activity was performed only once". Although acknowledging the limitations inherent in the study, the author still relies heavily on statistical significance testing without conducting any effect size analyses (cf. Table 6).

Although it is the submission policy of *Language Learning and Technology* (cf. Appendix B) for authors to follow the APA guidelines when submitting articles for publication, both of these studies ignored the requirement to conduct effect size analyses. Both of these studies relied heavily only on statistical significance testing to either confirm or reject the null hypotheses formulated for the studies. The authors, however, did acknowledge the limitations of their studies and did not attempt to generalise the findings of their research.

Table 6: Studies in Language Learning and Technology

Authors	Al-Seghayer (2001)
Research questions and/or hypotheses	This study investigates which of the image modalities – dynamic video or still picture – is more effective in aiding vocabulary acquisition. A null hypothesis of no difference was tested. Directional hypothesis: video clip annotations are more effective than still picture annotations in the acquisition of new words in a foreign language.
Participants	A convenience sample was selected. 30 ESL participants (17 males, 13 females) who were enrolled in the English Language Institute at the University of Pittsburgh. They are grouped according to their native language as follows: 13 Arabic, 4 Japanese, 6 Korean, 3 Spanish, and 4 Thai speakers. They had attained intermediate-proficiency TOEFL scores of 450-500.
Design	A within-subject design was employed (each participant served as his/her own control), allowing the same participants to be exposed to the same treatment conditions and then assessing them on the dependent variable after they participated in the experimental treatment. They were asked to read the story included in the multimedia programme, to take a vocabulary test, to respond to a questionnaire, and participate in a short interview. The independent variables, that is, printed text definition alone, printed text definition coupled with still pictures, and printed text definition coupled with video clips, were experimentally manipulated by exposing all participants to the experimental treatment condition. The dependent variable, that is, acquiring new words in a foreign language, was measured by two types of vocabulary tests: recognition and production.
Data analysis techniques and statistics reported	The Friedman test was employed to see if there were significant differences between printed text definitions coupled with video clips, printed text definitions coupled with still pictures, and printed text definition alone. Wilcoxon matched-pair signed-rank tests Bonferroni multiple-comparison procedure Percentages Statistical significance

Intervention	The interactive multimedia computer programme used in this study was designed by the researcher to enhance L2 vocabulary acquisition by providing readers with the meaning of a target word via hypermedia links to multiple modalities. The computer programme provided students who were reading a narrative English text with annotations for target words in various modes--text, graphics, video and sound all of which were intended to aid in the understanding and learning of unknown words. Instead of a traditional gloss, the programme provided various multimedia links. Users could read the target word's printed textual definition, hear its pronunciation, or view its meaning via a still picture or video. Participants met individually with the researcher at a computer lab where each was asked to fill out a background questionnaire and then given a brief introduction to the programme, its objectives, and its methods. The investigator demonstrated to each participant how the programme worked, that is, how to move from one section of the story to another, and how to click on an annotated word. The investigator also showed each participant that clicking on words allowed them to hear the selected word pronounced, read a definition, and see either a picture or a video clip. They were told that they could consult the annotated words whenever they wished and as many times as they wished. Each participant read the story individually, using the multimedia programme which contained multimodal annotations for the more difficult vocabulary items. Participants spent approximately 30 minutes reading the story. After reading the story, participants were asked to take a vocabulary test (described in the next section). Participants spent approximately 15 minutes completing the test. Participants were asked to respond to a questionnaire asking them to indicate which type of annotation mode helped them learn the annotated words best. Finally, participants engaged in a short interview in which they were asked to indicate whether the particular mode conveyed the

	meaning of the lexical item.
Control	-
Results	The results indicated that performance on the text-plus-video was significantly better than on the text-plus-picture condition ($z= 4.018$, $p< 0.001$), indicating that the video condition is more effective than the picture condition. It was also found that the performance on the text-only condition, and the text-plus-video condition was significantly different ($z= 4.219$, $p< 0.001$), showing that the former is more effective than the latter. However, there was no significant difference between performance on the text-plus-picture condition and text alone treatment in the printed text definition coupled with picture condition and the text definition alone condition ($z= 2.038$, $p= 0.042$). The qualitative data were collected by means of a questionnaire and interviews. The second instrument employed to determine the most effective annotation modes was a questionnaire. The results indicated that in the case of the video condition, 86.6% of the participants rated video as very helpful while 70% rated pictures as a very helpful mode. By contrast, only 10% of the participants said the textual definitions were very helpful. In fact, 60% said the textual condition was the least helpful as compared to 10% and 3.3% for the picture and video respectively. The last instrument employed to determine the most effective annotation modes was a face-to face interview. Participants agreed (90%) that the video clips showed the meaning of a word more clearly than other links. Almost all participants agreed that both the video (90%) and the still picture (76%) links showed what the words mean. However, for the textual definitions, they gave different opinions. About 46.6% agreed, 30% disagreed, and 23.33% were undecided.
Authors	Nikolova (2002)
Research questions and/or hypotheses	• If time on task is disregarded, acquisition of L2 words presented in a text via a multimedia instructional module is better when the students help create the instructional module rather than when they study the text containing the target words annotated with multimedia annotations by a teacher. • If time on task is taken into account as an additional variable, acquisition of L2 words presented in a text via a multimedia

	instructional module is better when the students help create the instructional module rather than when they study the text containing the target words annotated with multimedia annotations by a teacher.
Participants	The target population for subject recruitment was all students from the second semester, first-year French class at a large research university in the Midwest. The sample of subjects participating in the study was formed by all the students who volunteered to take part. Out of a target population of 69 students, 62 participated in the study, thus forming one experimental and one control group of 31 each. All students were native speakers of English.
Design	This study followed a randomized control-group pretest-posttest design. All subjects in the study were first randomly assigned to two groups. Each group was then assigned at random to either the control or experimental treatment.
Data analysis techniques and statistics reported	Descriptive statistics (e.g., means, standard deviations) MANOVA Wilks' Lambda F-value, df and statistical significance
Intervention	The experimental group had the same text as the control group. The text was presented via the same computer template, SmarTText. However, in the experimental setting, the target words in the module were not "hot"; they were not linked yet to their corresponding annotations. The experimental subjects had to link the target words with sound and picture files previously input in the computer by the researcher. For this purpose, they used the author mode of the SmarTText software. In addition, they had to write text annotations after looking up the meaning of the target words in a French- English dictionary and link these annotations to the target words. Both the experimental and control subjects were encouraged to ask for help if needed. There were quite a few questions asked in the experimental group. Most of the questions were related to the context in which the target words were presented, but there were also questions pertaining to the software. All subjects worked with headphones throughout the entire experiment. Therefore, the questions asked and the answers given could be heard only by the researcher and the individual who asked the particular question. At the end of the experiment, each experimental subject had to signal the researcher and his/her time out was logged in. In addition, the researcher examined the module produced by the subject. Despite several minor

	problems with the software, which required help from the researcher, all experimental subjects finished their modules and created correct hyperlinks for the annotations of the target words.
Control	The control group had to study the text in a multimedia module presented via computer. In this module the target words were "hot"; they were annotated by the researcher with text definitions (all 20 words), sound (all 20 words), and pictures (14 out of the 20 words) and appeared in boldface on the computer screen. By double -clicking on the "hot" word the subjects were able to see the text annotations (translation of the word). When a "hot" word was double -clicked in addition to the presentation of text annotations, which each word had, one or two of the icons in the right-hand column of the screen turned dark. Each dark icon indicated a link to a particular type of annotation -- sound or picture. All words had sound and text annotations, but only 14 out of the 20 words had picture annotations as well. For the word *décapotable* both the picture icon and the sound icon are dark. The word has, therefore, both picture and sound annotations. By double -clicking on the picture icon the students could see the picture annotation of the word. By double -clicking on the sound icon the students could hear a native speaker pronounce the word. The students of the control group were allowed to spend as much time as they desired studying the text and using the annotations. They were instructed to ask for help if they needed any. Only two students asked questions pertaining to the context in which the target words were presented.
Results	The study found that students have significantly higher rates of acquisition of L2 vocabulary if they participate in the authoring of a multimedia module than if they study one prepared by a teacher. The means of immediate vocabulary gain of the control group and the experimental group in the study were 13.00 and 14.94 words, respectively. The means of delayed vocabulary gain were 2.96 and 4.25 words, respectively. In conclusion, the first hypothesis was accepted and the second rejected. If time on task was not taken into account, the experimental subjects acquired vocabulary significantly better. If time on task was considered, there was no statistically significant difference between the groups as far as their vocabulary acquisition was concerned.

6.2.1.3 *Studies in Scientific Studies of Reading*

The methodology utilised by Symons et al. (2001) in both study one and study two was discussed in great detail (cf. Table 7). The intervention process followed with both the experimental and control groups was also described in detail by the authors. In these studies the authors used parametric statistics to analyse their data and report both statistical significance as well as effect sizes. Effect size results were presented as standard deviation units (cf. results section in Table 7) and not as either eta- or f-squared values (cf. Chapter 7). The conclusion reached by the authors is, therefore, justified, namely that: "This first study provides evidence that elementary students' independent information-seeking skills improve as a result of instruction in a strategic approach (Symons et al., 2001:16). The results of the second study indicated that providing students with a brief overview of the structure of informational books did not result in improved search success. Performance of students who were not instructed in strategy use suggests that children in the elementary grades do not spontaneously use an efficient approach to locating information in text. In conclusion, the authors state the following: "The direct and effective approach that was used in the current studies could readily be used by **actual teachers in real classrooms**" (my emphasis) (Symons et al., 2001:27). In the second study, the authors once again state that the effect size is 0.81, but they do not indicate in any of the tables where this value comes from and how it is calculated (cf. Study two results section, Table 7).

The study conducted by Castiglioni-Spalten and Ehri (2003) has several strengths (cf. Table 7). The data emerged from an experiment in which children were randomly assigned to conditions. Findings were replicated at two different test points. The post-tests were administered by a researcher who was not aware of children's treatment experience, thus reducing the influence of experimenter bias on outcomes. In addition to calculating statistical significance, the authors also calculated Cohen's effect sizes to assess the strength of the treatments' impact on outcomes. However, the study has limitations as well. The sample was relatively small even though it was sufficient to detect effects of the treatments statistically. Instruction was

limited to six sessions each lasting 20-30 minutes. Although this was sufficient to teach phonemic awareness, children might have attained mastery levels with more instruction. The authors conclude by stating that: "Although our study was a laboratory experiment conducted with individual students, it is likely that findings apply to teachers and classrooms" (Castiglioni-Spalten & Ehri, 2003:49).

Nunes et al. (2003:303) state that: "The results of our project are a step toward providing the basis for an effective bridge between research on children's reading and the study of methods of teaching children about morphological and phonological structures. The fact that **our interventions did not work as consistently** as we predicted should not be a discouragement" (my emphasis). The authors make these conclusions based on only the statistical significance of their results. No effect sizes were calculated in this study (cf. Table 7). The authors also used LSD tests (cf. results section, Table 7) that do not adjust for familywise error rate as opposed to Tukey tests.

In the study conducted by Christensen and Bowey (2005), the methodology section is discussed in great detail. Participants were randomly assigned to treatment conditions and the data was analysed using a multivariate analysis of variance as well as two-way analyses of variance. Although eta^2 was calculated and reported in the tables, the authors did not refer to these statistics in the reporting of their results or in the discussion section (cf. Table 7). The authors conclude by stating: "In summary, this study showed that programs designed to provide explicit practice in the use of letter-sound relationships to decode unfamiliar words can significantly enhance children's performance across a wide range of measures of reading and spelling" (Christensen & Bowey, 2005:347).

The methodological accuracy with which the authors, publishing in *Scientific Studies of Reading*, report and analyse their data seems to be higher than the studies reviewed in the previous journals. However, even though the authors tend to follow the recommendations of the APA in terms of reporting effect

sizes, very few of the authors actually address and interpret these results. It seems as if the calculations are there merely for appearance sake (e.g., Christensen & Bowey, 2005).

Table 7: Studies in Scientific Studies of Reading

Authors	Symons et al. (2001) STUDY ONE
Research questions and/or hypotheses	The purpose was to determine whether there were separate effects of being strategic in selecting text and of monitoring how well the extracted information fulfilled the search goal. Children in all grades would provide more accurate answers to questions when they were encouraged to monitor how well the extracted information fulfilled the search goal.
Participants	There were 180 students in Grades 3, 4, and 5 who participated in this study. The average age of the Grade 3 students was 8 years 8.5 months; Grade 4 students 9 years 10 months; and Grade 5 students 10 years 10months. The students were all attending public schools in a rural region of Canada.
Design	Students were randomly assigned to one of three strategy conditions within each grade, with 20 students in each condition in each grade.
Data analysis techniques and statistics reported	Accuracy, search time, and search sequence scores were analyzed using a 3 (grade) × 3 (strategy condition) **multivariate analysis of variance (MANOVA)**, yielding a significant effect of strategy condition, $F(6, 294) = 26.44$, $p < .001$, grade, $F(6, 294) = 9.280$, $p < .001$, and a significant Strategy Condition × Grade interaction, $F(12, 389) = 1.81$, $p = .044$. These significant overall effects followed up with **univariate analyses of variance (ANOVAs)** for each of the three main dependent variables. Specific main effects and interactions were then followed by **Tukey's post hoc tests** of means.
Intervention	**SLI** SLI involved the following steps: (a) finding words in a question to look up in the index, (b) using the index to locate pages of text to search, and (c) reading selected text page(s) carefully to identify the targeted information. **Strategy for locating information + monitoring (SLI + M).** Students in this condition were taught to use SLI and to monitor how well the extracted information fulfilled the search goal. The teacher modeled monitoring. In both the SLI and SLI+M conditions, the teacher reviewed the steps of the strategy.

	Student practice (SLI and SLI + M). Students practised locating answers to three questions. Practice questions were presented in counterbalanced random order. Children were also given a worksheet that prompted them to use the strategy. While the students practised using the strategy that they had been taught, the teacher answered their questions and gave guidance when appropriate to ensure that the students used the strategy.
Control	Students in this group were not given instruction of how to locate information in a book but completed the Wright brothers' question that the teacher used to model SLI and the three practice questions. They were not instructed in strategy use and were not given feedback on the adequacy of their answers to the search questions.
Results	**Accuracy** There was a significant main effect of strategy condition, $F(2, 171) = 20.88$, $p < .001$, grade, $F(2, 171) = 28.37$, $p < .001$, and a significant Grade × Condition interaction, $F(4, 171) = 2.56$, $p = .04$. SLI +M students (M= 2.5, SD = 0.8) found more correct answers to questions than students in the SLI condition (M= 2.0, SD = 1.0), who found more correct answers than the control group (M= 1.5, SD = 1.2). Grade 5 students (M= 2.6, SD = 0.6) were more successful at finding the information than were Grade 4 students (M = 1.9, SD = 1.0), who found more correct answers than Grade 3 students (M = 1.4, SD = 1.2). To examine the interaction, one-way ANOVAs were conducted at each grade with strategy condition as the between- groups factor and number of correct answers as the dependent variable. At Grade 3, this ANOVA was significant, $F(2, 57) = 15.22$, $p < .0001$. Tukey comparisons indicated that at Grade 3, both strategy groups found more correct answers than the control group, and the strategy groups were not different from each other. At Grade 4, the ANOVA was statistically reliable, $F(2, 57) = 4.45$, $p = .016$; students in the SLI + M condition found more correct answers than students in the control group. The SLI group did not differ from either the SLI + M or control groups. The ANOVA with Grade 5 students was also statistically significant, $F(2, 57) = 3.53$, $p = .036$, and again, SLI +M students found more correct answers than control students, but SLI was not different from the control group or from SLI +M. **The effect size with Grade 3 students** of the SLI + M instruction on accuracy (relative to the variability of the non-instructed controls) **was 2.04 *SD* units**. Whereas the effects of SLI +M on accuracy were smaller with older children (ES = 0.78

SD for Grade 4 and 0.56 *SD* for Grade 5), these are still of a moderate size, indicating that children benefitted from direct explanation and guided practice of a strategic approach to locating information, particularly when success monitoring was incorporated into the instruction. These smaller effects with older children may be due to the fact that the older children were successful without instruction, as evidenced by performance of the Grades 4 and 5 children in the non-instructed control group.

Searching Time

There was a significant main effect of strategy condition, $F(2, 149) = 9.32$, $p = .001$, and grade, $F(2, 149) = 20.85$, $p = .001$, but no Strategy Condition × Grade interaction, $F(4, 149) = 1.33$, $p = .26$. Tukey's honestly significant difference (HSD) comparisons indicated that SLI +M students ($M = 6.36$, $SD = 2.74$) and SLI students ($M = 6.98$, $SD = 3.12$) were faster than control students ($M = 8.94$, $SD = 4.10$). Grade 3 students ($M = 9.65$, $SD = 3.46$) spent more time searching the text than either Grade 4 students ($M = 7.13$, $SD = 3.31$) or Grade 5 students ($M = 5.81$, $SD = 2.77$). Strategy instruction reduced the amount of time that students spent searching the text, regardless of whether the students had been instructed to monitor the accuracy of their answers.

Sequence Data

A 3 (grade) × 3 (strategy condition) ANOVA with sequence scores as the dependent variable indicated a main effect of condition, $F(2, 171) = 86.921$, $p = .001$, and grade, $F(2, 171)$, 11.19, $p < .0001$, with no Condition × Grade interaction, $F(4, 171) = 1.43$, $p = .224$. SLI + M and SLI students had higher sequence scores than students in the control group. Sequence scores of SLI + M and SLI students did not differ. Grade 5 and Grade 4 students had higher sequence scores than Grade 3 students.

Process Measures

A number of comparisons were made on measures that indicated how students completed the search tasks. A 3 (strategy condition) × 3 (grade) ANOVA on the number of pages searched indicates a main effect of strategy condition, $F(2, 171) = 29.15$, $p < .01$, and grade, $F(2, 171) = 3.04$, $p = .05$, with no Strategy Condition × Grade interaction, $F(4, 171) = 2.33$, $p = .06$. Students in the SLI +M ($M = 5.6$, $SD = 1.9$) and SLI ($M = 6.7$, $SD = 8.8$) conditions searched far fewer pages of text than students in the control group ($M = 58.2$, $SD = 9.8$), $F(2, 171) = 29.15$, $p < .001$. The same analyses with number of choices made by the children as they searched the text indicated a main effect of condition, $F(2, 171) = 10.908$, $p < .001$, with both SLI and SLI + M students making fewer choices than did the control students. There were no grade differences, $F(2, 171) = 1.141$, $p = .322$, and no Grade

	× Condition interaction, $F(4, 171) = 0.067$, $p = .992$. The SLI strategy taught children how to use the index; 119 of the 120 SLI + M and SLI students used the index for each question, whereas only 16 of the 60 control children used the index. Students who used the index at least once per question ($n = 135$) found a mean of 2.3 ($SD = 0.9$) correct answers. The 45 students who did not use the index at all found a mean of 1.1 ($SD = 1.1$) correct answers. These means are significantly different, using a t test adjusted for unequal variances, $t(63.1) = 5.99$, $p < .001$.
Authors	Symons et al. (2001) STUDY TWO
Research questions and/or hypotheses	To assess whether an orientation to the structure of informational books would be as effective as strategy instruction.
Participants	Fifty-seven children from a rural elementary school participated in this study. The sample included 28 girls and 29 boys. Twenty-four of these children were completing Grade 3, and 33 were in Grade 4. The mean age of the students was 9 years 5 months ($SD = 6.6$ months).
Design	Students were randomly assigned to one of three instructional conditions (SLI + M, Textbook Organization, or Control) with 19 students in each condition.
Data analysis techniques and statistics reported	One-way ANOVAs with condition as the independent variable were followed up with Tukey HSD contrasts. Descriptive statistics (Mean, Standard deviation) Statistical significance Effect size stated, but how it was calculated is not indicated (e.g., eta- or f-squared?)
Intervention	Each child worked individually with an instructor for approximately 30 min. Children were introduced to the task in the same way as in the first study. SLI + M. Instruction in this condition was the same as in Study 1. Book structure overview. The teacher described to each student the structure of a traditional informational book, with emphasis placed on the distinction between a table of contents and an index.

Control	Each student was given 2 min to scan the book prior to completing the practice questions. They then completed the practice questions unassisted.
Results	**Search Accuracy** There was a significant main effect of condition, $F(2, 51) = 3.23$, $p = .048$, and grade, $F(1, 51) = 4.19$, $p = .046$, and no Condition × Grade interaction, $F(2, 51) = 0.62$, $p = .54$. The mean accuracy of children in the SLI + M condition ($M = 1.82$, $SD = 0.97$) was significantly higher than that of children in the control condition ($M = 0.95$, $SD = 1.07$). No other differences were statistically reliable. Grade 4 children ($M = 1.62$, $SD = 0.86$) provided more correct answers than Grade 3 children ($M = 1.08$, $SD = 1.10$). The **effect size** of the strategy with monitoring instruction (in comparison to the non-instructed control group) **was 0.81**. Providing students with a brief overview of the structure of informational books did not result in improved search success. **Search Time** There was a significant main effect of condition on time, $F(2, 47) = 13.17$, $p = .001$. Students in the SLI + M condition spent less time searching for the information ($M = 6.16$ min, $SD = 2.59$) than students in the control condition ($M = 11.82$ min, $SD = 3.30$). There was no effect of grade, $F(1, 47) = 0.12$, $p = .74$, and no Grade × Condition interaction, $F(2, 47) = .1.54$, $p = .23$. Grade 3 children spent as much time searching the text ($M = 8.77$ min, $SD = 3.29$) as the Grade 4 children ($M = 8.47$ min, $SD = 2.39$) but provided fewer correct answers. **Sequence Data** Instructional condition had a significant effect on the quality of the students' search sequences, $F(2, 51)=31.32$, $p=.001$. Search sequences of children in the SLI condition ($M=18.14$, $SD=1.80$) were significantly higher than those of the children in the control condition ($M=6.68$, $SD=5.67$). Again, the mean of the structure overview group did not differ reliably from either the SLI + M or control group.
Authors	Castiglioni-Spalten & Ehri (2003)
Research questions	• Novices who are taught phonemic segmentation will outperform untaught novices in learning to read words by memory and in inventing spellings using partial alphabetic cues.

and/or hypotheses	• Novices taught to segment by attending to articulatory gestures will read and spell better than children taught to segment without gestures. • Segmentation training will not help children decode novel words because segmentation instruction does not teach blending.
Participants	Participants were selected from two urban schools located in lower socioeconomic status neighborhoods. These schools did not teach children to read in kindergarten. There were 75 students who returned permission forms and were screened for the study. To be included, children had to meet the following criteria that placed them in the partial alphabetic phase with limited PA but proficiency in English: name at least 13 of the 17 target letters, segment at least one phoneme within 1 of the 10 words correctly but segment no more than 3 words correctly, read no more than 1 consonant–vowel–consonant (CVC) non-word and no more than 10 primer-level words, identify at least 35 vocabulary words correctly on the Peabody Picture Vocabulary Test–Revised (PPVT–R; Dunn & Dunn, 1981), and perform similarly enough to two other children to form a triplet. There were 45 children selected (29 girls, 16 boys; *M* age = 5 years 9 months). They varied in the language spoken at home: 30 English, 12 Spanish, and 3 another non-English language. None was enrolled in an English as a Second Language class.
Design	Pre-test post-test experimental (2 groups) control group design
Data analysis techniques and statistics reported	Descriptive statistics (Mean, Standard deviation) ANOVA ($F(2,39)=11.14$, $P<0.001$) Tukey's pairwise comparison procedure Statistical significance Cohen's effect sizes
Intervention	**Mouth condition.** First children learned to associate their mouth movements with pictures and sounds. The experimenter explained and illustrated how the mouth moves in various ways to pronounce words, children examined their mouths in a mirror, and they pointed to their own articulators. Then four sets of pictures were taught. Each set consisted of four pictures showing the mouth articulating

different sounds:

- Set 1: /p/, /t/, /f/, and /long a/
- Set 2: /b/, /v/, /d/, and /long e/
- Set 3: /m/, /l/, /k/, and /s/
- Set 4: /z/, /n/, /g/, and /long o/

The experimenter explained each picture in a set and then had children practice listening to randomly ordered sounds and pointing to the corresponding pictures until succeeding on eight in a row. To verify their answers, students were taught to look in a mirror at their own mouths pronouncing the sounds. Children all succeeded in learning all four sets of pictures to criterion. After children had learned the correspondences, they began segmentation instruction. They were provided with a horizontal row of empty squares to be filled with picture blocks representing each sound in the words. Children were informed how many blocks and sounds to detect by the number of squares displayed either two, three, or four. Placed behind the row of squares was a large mirror for children to use in checking their mouth positions as they pronounced and segmented words. The experimenter showed them how to perform the segmentation response. The procedure included the following steps:

1. The experimenter pronounced the word and showed its picture (e.g., *ape* and picture of the animal). The child repeated the word.
2. The child identified the sequence of phonemes in that word by pronouncing each separately (e.g., /e/–/p/) while selecting its mouth picture from an array of eight picture blocks and placing the block in an empty space sequenced from left to right. If the child segmented incorrectly, the experimenter provided corrective feedback. Children began by selecting blocks to fill two squares representing two-phoneme words—for example, *go*, *ape*, and *sea*. When they had correctly segmented eight in a row, they proceeded to CVC words segmented into three squares—for example, *bake*, *leaf*, and *soap*. Practice continued until they correctly segmented eight in a row. Then they progressed to words with consonant clusters —for example, *flea*, *place*, and *beast*. Practice continued until they correctly segmented eight in a row or until they had received six instructional sessions.

Ear condition

First, children learned to associate blocks with a succession of sounds and to distinguish whether the sounds were the same or

	different. The experimenter referred to an illustration of the inner ear chamber to explain how the ear works to perceive sounds. Then she introduced blocks each depicting a drawing of an ear. She taught students to select and tap blocks on the table to show how many separate sounds were heard and whether they were the same or different. For example, if the experimenter said "/p/, /p/," the child picked up one block and tapped it twice on the table while repeating the two sounds. If she said "/p/, /t/, /f/," the child selected three different blocks and tapped each block once on the table while pronouncing each sound. Practice began with two sounds, then progressed to three sounds, and then four sounds. In each case, the child practiced these responses until reaching a criterion of two correct in a row. No one failed to reach criterion. The segmentation instruction then commenced. This instruction was identical to the mouth condition except that children segmented with blocks lacking articulatory pictures. They pronounced the sounds in words and placed a block for each sound in a horizontal row of empty squares showing the number of sounds to be detected. No mirror was present. The same words and non-words illustrated on cards were presented. Children practiced each list of two-phoneme, three-phoneme, and consonant cluster words to the same criterion of eight successive items correct or a maximum of six sessions.
Control	Children in the control group remained in their classrooms between the pretest and posttests. They received regular kindergarten literacy instruction that included memorizing and writing alphabet letters and singing the alphabet song. Neither phonemic awareness nor formal reading instruction was taught in these kindergartens.
Results	The reliability of posttests administered 1 week apart was assessed with Cronbach's alpha. Results revealed high reliabilities: segmenting words, 0.89; segmenting sounds, 0.88; spelling sounds in words, 0.92; learning to read words correctly, 0.80. The reliability of the non-word reading posttest was not calculated because most scores were zero. To assess effects of PA instruction on outcomes, ANOVAs were conducted with treatment (three conditions) and time of posttest (immediate vs. delayed) as the independent variables. Treatment main effects were followed up by post hoc tests using Tukey's pairwise comparison procedure. Effect sizes were calculated also to assess the strength of the treatments' impact on outcomes. According to Cohen's (1988) rule of thumb, an effect size of 0.20 is considered small, 0.50 is moderate, and 0.80 is large.

	Three students were unavailable for the delayed posttest—two in the mouth condition and one in the ear condition. ANOVAs were conducted in two ways to determine effects of the three missing students: (a) ANOVAs that included all remaining students (n = 42) and (b) ANOVAs that included only the 13 intact triplets (n = 39). Results of the ANOVAs were the same except for one post hoc Tukey test1 involving the spelling task. We report results involving the larger sample in the following discussion. To assess whether PA instruction was effective, two segmentation posttests were administered—one at the end of instruction and one after a week's delay. This task differed somewhat from the task used to teach PA in that no markers were used, and students were not told how many segments to find in words. Two ANOVAs were conducted—one on the number of individual phonemes segmented correctly and one on the number of words segmented correctly. Main effects of treatment group were statistically significant in both analyses. Neither main effects nor interactions involving time of posttest were significant, indicating that little forgetting occurred 1 week after training (Fs < 1). Mean performance is reported in Table 2 where effects across both posttests are combined. Tukey post hoc tests revealed that the mouth and the ear groups each statistically outperformed the control group in segmenting phonemes as well as words correctly but did not differ from each other. Effect sizes were very large—1.47 (mouth) and 1.53 (ear). Thus, both types of PA instruction proved effective in teaching children to segment. The words that students segmented varied in complexity. Students who received PA instruction segmented, on average, 62% of the CV and VC words correctly, 28% of the CVC words, and 11% of the words containing consonant clusters. No child in the control group segmented any consonant cluster words correctly. These findings show that, although PA-trained children improved substantially in their ability to segment, instruction did not produce mastery levels of performance, particularly on words with consonant clusters. One reason may be that changes in segmentation procedures on the posttest may have depressed performance. Nevertheless, instruction was highly effective in boosting performance of both treatment groups above that of the control group as indicated by the very large effect sizes. We expected PA instruction to transfer and help children read and spell even though it did not teach students to manipulate letters. This was because students could name most of the letters whose sounds they learned to segment, most of these letters contained the relevant phonemes in their names, and students were able to produce semi-phonetic spellings on the pretest. The spelling posttest required students to write letters for the sounds in words.

	In the ANOVA, the main effect of treatment group was significant, but neither the main effect nor the interaction involving posttest time was significant (*F*s < 1). A post hoc Tukey test revealed that both the mouth and the ear groups spelled significantly more sounds than the control group but did not differ from each other. Effect sizes were large. These findings indicate that both types of PA instruction facilitated students' ability to spell sounds in words. Results showed that both the articulatory and the ear methods taught PA effectively compared to the no-treatment control group. This skill transferred to a spelling task for both groups who wrote phonemes in words more accurately than the control group. However, only articulatory training transferred to a reading task. The hypothesis that both the mouth and the ear methods would teach phonemic segmentation effectively was supported. However, children did not display mastery on the posttests. Contrary to our hypothesis that articulatory training would teach segmentation more effectively, the two forms of PA instruction proved equally effective when sufficient time for training was provided.

Authors	Nunes et al. (2003)
Research questions and/or hypotheses	• Intervention that concentrates exclusively on children's awareness of the morphemic structure of spoken and written words will improve children's learning of morphologically based spelling rules. • Phonological interventions will help children with phonologically based rules but should not affect their learning, for example, how to spell the ending of words like *emotion*.
Participants	The children were recruited from eight London state schools; four schools provided the experimental groups, and four provided the control group. The experimental groups included 222 children (112 boys and 110 girls), and the control group included 246 children (121 girls and 125 boys). Six schools contained children from a wide range of socioeconomic backgrounds. In the other two schools the children were mainly from middle-class backgrounds. Those from one of these two schools were assigned to the experimental groups and from the other to the control group. Each of the six other schools was assigned to the control or the experimental groups. In each class, the children were randomly assigned to one of the four intervention groups, with the restriction that at least two children of the same sex were part of each group. In the intervention sessions, these children were seen in groups varying in number from 5 to 8. The project included 32 sets of children, 4 sets in Years 3 and in Year 4 in each of the four schools.
Design	Pre-test post-test experimental and control group design
Data analysis techniques and statistics reported	One-way analyses of covariance Post hoc least significant pairwise comparisons Statistical significance Descriptive statistics (Mean, Standard Error)
Intervention	Small groups of 4 to 8 children were taught in 12 weekly intervention sessions, which consisted of group games. The procedures were purely oral for children in the morphological training alone and phonological training alone groups. The children in the *with writing* groups were taught to instantiate the rules in writing.

	Morphological training The morphological groups were taught about word stems and grammatical categories in relation to inflectional affixes and derivational affixes. In the morphology training with writing group, where the intervention was connected with writing, the children were instructed on how to use this information to produce spellings. Sentences like "A person who does magic is a what?" to introduce the children to agentives. The children completed the sentence orally and wrote the word magic plus the ending that made it into magician. They also learned how to use this information to analyse long words that they wanted to decode. In the morphology training alone group, the children identified the agent and completed the sentence but did not write down the words. **Phonological training** The phonological groups were taught mainly about more difficult phonological awareness tasks and long and short vowels combined with a variety of operations (e.g., blending). The instruction contained games that could be responded to orally and in writing. In the phonology training alone group, where instruction was only oral, the children would produce oral answers. In the phonology training with writing group, where instruction was combined with writing, the children had a model of both rimes on cards in front of them and had to say and write their answers. In this group the children were also taught how to use information from these phonological rules in decoding words.
Control	The control group received similar teaching in the classroom.
Results	Intervention led to higher performance by all four intervention groups (in comparison to the control group) in a standardized test of reading. The group effect was significant in the analysis of the standardized reading scores, $F(4, 394)=6.46$, $p<0.001$. Post hoc **least significant pairwise comparisons** established that all four intervention groups significantly outperformed the control group ($p<0.001$) for the phonology training alone group and $p<0.01$ for the other three intervention groups. Morphological intervention affected progress in the use of morphological rules in spelling The effects of intervention on the use of morphological rules in reading were weaker and less specific. Intervention had no discernible effect on the use of phonologically based conditional spelling rules.

Authors	Christensen & Bowey (2005)
Research questions and/or hypotheses	The study examined the efficacy of three approaches to teaching reading skills. Two of the programmes explicitly taught decoding skills. One focused on orthographic rimes (using word families) and one focused on grapheme–phoneme correspondences. The third approach acted as a treatment control. It used an implicit phonics approach by teaching letter-sound correspondences within a whole language approach to reading text.
Participants	Participants were 116 children in their 2nd year of schooling from 7 classrooms in two state elementary schools in Brisbane, Australia. The 1st year of schooling in the state of Queensland is termed Grade 1. Thus, these children were in their 2nd year of schooling, Grade 2. The average age of the children at the commencement of the study was 7 years 1.5 months (*SD* = 3.17 months).
Design	Randomised pre-test post-test experimental and control group design.
Data analysis techniques and statistics reported	Descriptive statistics (e.g., means, standard deviations) MANOVA Two-way analyses of Variance Tukey Eta^2 Statistical significance
Intervention	There were three instructional conditions. Two explicit decoding conditions provided teaching and practice in the analysis and synthesis of words using sub-lexical units. The **Orthographic Rime programme (OR programme)** focused on orthographic rime correspondences and onset-rime blending, and the **Grapheme–Phoneme Correspondence programme (GPC programme)** focused on individual grapheme–phoneme correspondences and the analysis and synthesis of individual phonemes. The third programme acted as a control. It taught sound–symbol relationships using an implicit approach to phonics. This taught decoding skills using symbol–sound relationships as they occurred when reading authentic text. All classrooms used a whole language approach to reading in their normal lessons. This included whole class reading of "big books" and individual

	reading of "authentic texts." There was a focus on reading a variety of genres, so children were given both fiction and informational texts to read and discuss as a class. They were encouraged to write using their own invented spellings. The implicit approach was consistent with the usual reading instruction in the classes. As far as possible, the three programmes corresponded in key instructional features. The programmes were organized into modules. Children in the two explicit decoding instruction programmes practiced the same number of words in each session and the same words across each module of 10 lessons. Modules consisted of 8 lessons each introducing a set of four new words. In addition there were 2 review lessons. Children in the skills-based groups worked with one list of words per day. However, the words within sets for each programme were presented in different orders and combinations, so that the primary instructional focus was on either orthographic rimes or grapheme–phoneme correspondences. New letters were introduced at the same rate and as close as possible to the same order. All three programmes introduced the same limited number of letters in each session. Each list in the OR programme contained words with the same orthographic rime or word family. The GPC programme differed from the OR programme in teaching segmentation and blending skills by focusing on manipulation and representation of phonemes rather than orthographic rimes. Lists in the GPC programme did not contain any rhyming words. For example, one OR list consisted of *top, mop, hop, shop*. The corresponding GPC list was *mat, hop, run, shin*. The programmes were presented in a series of six modules, each containing eight sets of four words. Each module was divided into lists of four words. Both programmes covered the same words in each module. There were review lists that included previously covered words midway through and at the end of each module. Review lists were confined to words in this module and did not cover words that had appeared in previous modules. In a sense, the GPC programme served as a "control" for the OR programme. The OR programme was constructed first. Sets of 4 new words from a single word family were devised and organised into modules of 32 words (eight sets of 4 words, with each set introducing a single orthographic rime). For instance, Word Set 7 of Module 4 of the OR programme contained the words *damp, ramp, stamp*, and *camp*. The 32 words within each module were then rearranged into eight new 4-word sets for the GPC programme. Because of the need to correspond to the OR programme, the words within each four-word set of the GPC programme were not necessarily optimal in terms of the grapheme–phoneme correspondences taught. Nevertheless, as far as possible, GPC programme word sets contained at least two words that forced fine-grained analyses of somewhat similar

	orthographic rimes. Thus, GPC programme Word Set 7 within Module 5 contained the words *coat, leak, drain,* and *snail*, with *drain* and *snail* forcing analysis of the two similar orthographic rimes, *ain* and *ail*, and reinforcing the *ai* correspondence. Many word sets contained either two such pairs of words or a group of three such words. Thus, GPC programme Word Set 5 in Module 4 contained the words *thump, stamp, kill*, and *smell*, and GPC programme Word Set 2 in Module 5 contained the words *feed, sweet, sheep*, and *trick*.
Control	The implicit phonics control programme covered the same letter-sound correspondences as the other programmes. This included all initial consonants, medial vowels and final consonants, and consonant blends. For example, when the final consonant cluster *mp* was introduced in the skills programme it was also introduced in the implicit phonics programme. However, where the skill programmes covered grapheme–phoneme correspondences in the context of reading isolated words, the implicit phonics programme introduced letter-sounds in the context of reading books. Thus, children were not given practice in decoding isolated words. Additionally, the constraint of locating texts that contained words with a specific set of letter- sounds meant that the implicit phonics programme did not necessarily contain precisely the same words as the other programmes. For example, when the skill-based programmes covered the letter-sound correspondence *oa,* the list contained the word *coat*. However, in the corresponding lesson the children in the implicit phonics programme encountered the word *boat*. Overall, the implicit phonics programme contained the same number or more exemplars of words containing each letter-sound correspondence as the skill-based programme. If children made an error when reading one of the target words, they were asked to look at the letter-sound they were learning and to see if they could "work out" the word. If they continued to make an error the teacher modeled the process of working out the word until the student could say the correct word. Children in the control programme worked with the teacher to read a piece of text from an age-appropriate storybook. This was followed by a discussion of the story. The letter-sound correspondences for the day were introduced on letter cards. The same letter-sound correspondences were introduced to all groups at the same time. For example, when the letter *s* was introduced to children in the skills programmes, it was also introduced to the implicit phonics group. As with the skills programmes, children thought of words that began and ended with the sound. The group then went through the story to locate words that contained target letter-sound correspondences. Children then

	completed workbooks that contained sentences from the story containing words using this letter-sound correspondence. In these workbooks, spaces for letters representing the target letters were left blank. Children completed the activity by inserting the appropriate letter in the spaces.
Results	Children in both explicit decoding programmes performed consistently better than the control group in the accuracy with which they read and spelled words covered in the programme. Only children in the grapheme–phoneme correspondence programme consistently spelled transfer words better than children in the control group. In addition, children in the grapheme–phoneme correspondence group consistently read words more quickly than children in the control group. Children in both explicit decoding programmes scored higher than the children in the control group on measures of reading comprehension and oral reading at posttest.

6.2.1.4 Studies in System

A potential weakness in the design of the Lao and Krashen (2000) study is that the comparison group was substantially less competent in English at the start of the experiment. It is also questionable whether the use of ANCOVA is appropriate for this experiment; experimental and comparison groups were not randomly assigned and the two groups may be very different; the experimental group consisted of translation majors, and the comparisons consisted of social science majors. The authors rely exclusively on the statistical significance of the results and do not calculate any effect size measures (cf. Table 8). It is seems as if the authors were working with two populations of students and the calculation of significance testing is, therefore, not relevant, but the calculation of effect size is very important. Even the study has serious methodological limitations; the authors make the following statement: "**We can safely conclude**, therefore, that reading for meaning is the factor that distinguishes the difference in gains between the two groups, either reading done in class, out of class, or both" (my emphasis) (Lao & Krashen, 2000: 268).

In the study conducted by Dreyer and Nel (2003), the authors report the statistical significance (P-value) as well as the practical significance (i.e., Cohen's effect size d) of the results (cf. Table 8). In reporting their results the authors state: "The post-test results, however, indicated that the students in the experimental group used certain strategies statistically ($P<0.05$), as well as practically significantly (small to large effect sizes), more often than the successful students" (Dreyer & Nel, 2003: 356). Based on the results, they come to the following conclusion: "The present findings suggest that students benefit from strategic reading instruction offered in a technologically-enhanced learning environment" (Dreyer & Nel, 2003: 362). In their analysis and interpretation of the data, the authors are far more justified in the conclusions they reach. However, one limitation of the study is the use of a quasi-experimental non-randomised control group design.

System does not require authors publishing in the journal to utilise the APA recommendations for reporting experimental research. A review of the articles

published in this journal, therefore, indicate that most authors still prefer reporting only the statistical significance of their results.

Table 8: Studies in System

Authors	Ying Lao & Krashen (2000)
Research questions and/or hypotheses	This study evaluates the impact of popular literature study on English as a foreign language reading competence and reading attitudes.
Participants	A total of six classes, 91 students, participated in the study as experimental group subjects. All students were first year university students majoring in translation. Comparison group students were 39 social science majors who were enrolled in a traditional academic skills development course that covered oral skills, writing, listening and reading. These students were also first year university students who had studied English for 15 years.
Design	Pre-test post-test comparison group design.
Data analysis techniques and statistics reported	Mean gain scores t-test, F-values statistical significance ANCOVA
Intervention	Students were required to read six books in one semester (14 weeks). Five out of six were assigned and one book was chosen by the students themselves based on their own interest. Self-selected books included *The Bridges of Madison County, The Gift, Circle of Friends, Jurassic Park*, and *Anne's House of Dreams*. Vocabulary size was estimated by counting the number of words on three pages, dividing by three, and multiplying by the number of print-filled pages in the book. The course consisted of four major components: 1. Students did the assigned reading before coming to class and read one book every 2 weeks. Five appealing as well as challenging novels were selected and were provided for the students as assigned readings. The course began with the simplest

	book (*Charlotte's Web*). Towards the end of the semester, students had the opportunity to select a book of their own choice. 2. Films/videos: for four of the books (all except *The Catcher in the Rye*), films or videos were available. Students saw the film or video version after they had read approximately half of the book. Seeing the film or video version of the story did not appear to diminish students' interest or make them less curious to read the rest of the book; in fact, it appeared to have the opposite effect. 3. The majority of class time was spent discussing the reading. To aid in the weekly discussion, students were asked to jot down their thoughts or questions about language and/or content that arose while they were reading and bring them to class for discussion. Class time also included explicit instruction on reading strategies, all intended to encourage students to read for meaning. Students were told not to look up all unknown words in the dictionary, but to utilize context clues and to read for general gist and not for details; students were told that they would not be tested on the details of content and that they should not try to remember what they were reading as they were reading; rather, they were free to enjoy the story. They were, in other words, told to read English the way they would read for pleasure in Chinese. At the beginning of the semester, students in all of the experimental classes requested direct instruction on vocabulary, including a vocabulary list with definitions and some class time devoted to discussion of difficult words. This was provided, but proved to be unpopular among students, according to their written comments at the end of the semester. 4. Students were asked to do two short (one to two page) essays. They could either discuss how the assigned novel related to their lives or write a continuation of the story. The final project was to review the book they had selected.
Control	Comparison group students were 39 social science majors who were enrolled in a traditional academic skills development course that covered oral skills (e.g. delivering academic or technical data in an oral presentation), writing (e.g. organization of essays, editing, proof-reading), listening (e.g. note-taking during lectures), and reading (e.g. taking notes from academic texts). All students were required to do a research project which was graded on content, organization, and the quality of the abstract and bibliography. Comparison group subjects were enrolled in two separate classes (20 students in one, 19 in the other), both taught in Spring 1997. Similar to the experimental group, these students were also first year university students who had studied English for 15 years and had graduated from high schools in which English was the medium of instruction. Experimental and comparison

	classes were taught by the same instructor (C.L.), who attempted to teach all sections with the same dedication and enthusiasm.
Results	Students who participated in a popular literature class designed to interest them in pleasure reading increased their reading accuracy and reading rate substantially, far more than the comparison group. Experimental post-test scores were significantly higher in both cases (for accuracy, $F=111,87$, $P<0.0001$; for rate, $F=233.75$, $P<0.0001$). Experimental subjects also indicated that they were more interested in pleasure reading as means of improving their English than they were before taking the class, and felt that the literature class would help them in future study. Comparison students were not as enthusiastic about their class. These results are consistent with those of studies showing the positive impact of free reading. The results of the Reading Attitude Survey are consistent with those of previous studies that show that reading is perceived to more pleasant than direct grammar instruction. Reading for meaning is the factor that distinguishes the difference in gains between the two groups, either reading done in class, out of class, or both.
Authors	Dreyer & Nel (2003)
Research questions and/or hypotheses	• What does the reading comprehension and reading strategy use profile of first-year students at Potchefstroom University look like? • Did the students in the experimental group who completed the strategic reading component of the English for Professional Purposes course in a technology-enhanced environment attain statistically and practically significantly higher mean scores on their end-of-semester English, Communication and TOEFL reading comprehension tests, and did they differ significantly in terms of their reading strategy use?
Participants	All first-year English as a Second Language (ESL) students (n=131) taking the English for Professional Purposes course participated in this study. The participants included speakers of Afrikaans and Setswana majoring in Communication Studies. The students were randomly allocated into the experimental or control groups. Within the experimental and control groups, the students were divided into two additional groups, namely successful and unsuccessful or "at risk" for failure. The students were divided into these two groups based on their scores for reading comprehension tests in English, Communication Studies and the TOEFL. All those students who obtained percentages below 55% were categorised as "at risk", whereas the students who

	obtained percentages above 55% were categorised as "successful".
Design	A *quasi-experimental non-randomised control group* design.
Data analysis techniques and statistics reported	Descriptive statistics (e.g., means, standard deviations) T-tests Cohen's effect size *d*
Intervention	Varsite is a Learning Content Management System (LCMS). A LCMS is a multi-user environment where lecturers can create, store, reuse, manage, and deliver digital learning content from a central object repository. A LCMS contains four basic elements: (1) a dynamic delivery interface (providing links to related sources of information, resources, the electronic study guide, and supports assessment with user feedback), (2) an automated authoring system (used to create the reusable learning objects that are accessible in the repository), (3) an administrative system (used to manage student records, track and report student progress, and provide other basic administrative functions), and (4) the learning object repository (serving as a central database in which learning content is stored and managed, and made accessible to the learners). The students had access to the following features within the Varsite environment: (1) electronic study guide, (2) announcement section, (3) assignment and resource section, (4) assessment section, and (5) interaction with peers and instructors. Each of these is described below. The electronic study guide differed from the printed interactive study guide in that it contained only the main points of emphasis on the reading process and the various reading strategies. It did not contain detailed explanations or examples. The purpose of the electronic study guide was to provide a quick reference for students while they were completing tasks that required them to follow a number of hyperlinks. For example, if the students wanted to know about text structure they could simply click on study guide link and they would be taken to the relevant page in the electronic study guide.

	The second feature was the announcement section. Here, the lecturers informed the students on a daily basis of assignments that had to be completed as well as due dates. In the assignment and resource section students were given a detailed outline of the tasks to be completed; the resource section contained two sub-sections, one on general topics and one specifically for Communication Studies. The resource section also contained a number of hyperlinks that were updated on a weekly basis to ensure that students had access to a plethora of information on the specific topics being discussed in their Communication Studies class. The English lecturers coordinated their teaching schedules with that of the Communication Studies lecturer. During the first 7 weeks of the semester, the lecturers provided the students with a variety of generic topics (e.g., current news, music, business reports, etc.), as well as a number of hyperlinks (i.e., scaffolding) that they had to use in order to gain access to the information needed for the completion of the tasks. During the last 6 weeks of the semester, the students were allowed to "surf" the Internet on their own, with only limited guidance from the lecturers, in order to find the information needed to complete the assignments. The assignments focussed on the use of reading strategies (e.g., predict what information the following website will contain; formulate a number of questions you want answered after reading an article on non-verbal communication, etc.). The fourth feature, the assessment section, was used in order to set a number of online practice assessments. Students had to make use of a variety of reading strategies in order to complete the assessments (e.g., identify the purpose of a selected piece of text, identify the main idea, make inferences, predict, formulate questions, summarise, etc.). The fifth feature was interaction with fellow students and also with the lecturers. This was accomplished via email. In general, the Varsite environment exposed students to a variety of authentic information that increased their background knowledge and comprehension of topics they were also discussing in their Communication Studies class (e.g., small groups, conflict in small groups, etc.). Some of the sites included video and audio clips (e.g., interviewing, negotiation skills, etc.). Initially, the activities and tasks were lecturer-guided, but as the students gained confidence, they were allowed to make their own choices. The rationale for using selected readings from the Internet was to surpass what the lecturers could offer in the contact sessions.

Control	**Interactive study guides** At Potchefstroom University printed interactive study guides are compulsory for all full time courses on campus. The authors of the strategic reading study guide tried to obtain a balance among three aspects: (1) the core information (i.e., the content on strategic reading), (2) the tasks and activities for learners to actively interact with the various sections of the module in order to develop the application of knowledge and skills in terms of the outcomes, and (3) encouragement of learners to manage their own learning. The major focus in the study guide was on explaining the main features of a particular strategy and explaining why that strategy should be learned (i.e., the potential benefits of use). The benefit of use was linked to students' reading profiles. In this way, students could see the necessity of reading strategy use, as well as the link to their reading comprehension ability. Appendix contains an outline of the content of the study guide, as well as the outcomes formulated for the strategic reading component. In the study guide, the following aspects formed a minor focus: (1) how to use the strategy, (2) when and where the strategy should be used, and (3) how to evaluate the use of the strategy). The study guide, therefore, contained sufficient explanation about strategic reading, but only a few practice activities. **Contact sessions** The purpose of the contact sessions was to give the students additional information on the strategies, to model the strategies for the students, and to provide practice opportunities both individually and in groups. During the first two sessions, the students were given information on the importance of motivation, anxiety, and time management because of the important role these variables play in language learning. In addition, the students and the lecturers brainstormed on reading strategies, and they discussed their prior experience with the use of reading strategies and the rationale for using them. At first, the discussion was linked to general topics (e.g., reading magazines, short stories, cookbooks, maps, etc.) and then specifically to content in their major (e.g., mass communication, non-verbal communication, communication theories, etc.). During the contact sessions, a brief overview was given of what a strategy is and why it should be used (i.e., minor focus). The major focus during the contact sessions was on how to use the strategies, when and where to use them, and how to evaluate their use. The authors tried to build from the student's understanding of whatever strategies he/she was currently using to placing these strategies in question by testing their validity against the task demands placed upon them by higher education. During the course of the 13-week

	semester, the students were given the opportunity to practice with simple sentences, then with paragraphs, then with a variety of genres, and lastly, with the content of their major (i.e., Communication Studies). Students were also shown how to set a purpose for their reading and how to approach the reading of different texts (e.g., narrative versus expository).
Results	An analysis of the reading comprehension scores (pretest) of the students in the experimental and control groups indicated that there was not a statistically significant difference in their mean scores on any of the reading comprehension measures. The language proficiency scores, as measured by the TOEFL, of the students in both groups ranged from 400 to 599. These scores indicate that some of the students' proficiency levels can be considered to be too low for academic work. A closer analysis of the TOEFL scores indicated that the at-risk students in this study achieved the lowest score in the reading section of the TOEFL test. This is a major cause for concern, especially when one considers that students need to read and comprehend a large number of academic texts. In the English for Professional Purposes course offered at Potchefstroom University, 30.53% of the students enrolled in this course were identified as being "at risk" for failure or unsuccessful. The mean pretest reading comprehension scores, on the English, Communication and the reading section of the TOEFL test, for the at-risk students were all below 55%. The results indicated that the at-risk students differed statistically ($P<0.0001$), as well as practically significantly ($d>0.8$), from the successful students on all the reading comprehension measures. In terms of reading strategy use (pretest), the results indicated that there was not a statistically significant or a practically significant difference in the reading strategies used by the students in the experimental and control groups. The posttest results, however, indicated that the students in the experimental group used certain strategies statistically ($P<0.05$), as well as practically significantly (small to large effect sizes), more often than the successful students. A closer analysis of the reading comprehension scores (posttest) of successful and at-risk students in the experimental and control groups indicated that the successful students in the experimental group as well as the at-risk students in the experimental group achieved statistically ($P<0.05$), as well as practically significantly (small to large effect sizes), higher mean scores on the reading comprehension measures in comparison to the successful students as well as the at risk students in the control group.

	The results indicated that students who received strategic reading instruction in a technology-enhanced learning environment received both statistically and practically significantly higher marks on three reading comprehension measures than did the students in the control group. This was true for successful students as well as for those considered to be at risk.

6.2.1.5 Studies in Teaching English as a Second or Foreign Language (TESL-EJ)

In the study conducted by Levine et al. (2000), the authors report both quantitative and qualitative findings indicating that many researchers are opting for a combination of the two methods in order to provide more enriched data and possible explanations for certain reading problems (cf. Table 9). Although the groups were randomly assigned to either the experimental or the control groups, the groups participated in the study in an intact format. The findings of this study indicate that the computerised reading environment **"may contribute"** (my emphasis) to the development of EFL critical literacy skills more than the conventional reading environment. This conclusion is based on the statistical significant difference that was found between the global reading skills of the experimental and control groups. No effect size calculations were done in this study. It should be noted, however, that the authors are far more cautious in the way that they phrase the presentation of their results (see bold type above).

The description of the methodology employed in Herman's (2003) study seems to be very limited. The author does not give sufficient detail in terms of the intervention that was used in the study. A quasi-experimental non-randomised design with intact classes constituting a very small sample size (n=17) was used in the study (cf. Table 9). The findings in the study were not as the author expected and he attributes this to the lack of sensitivity of the instrument that was used to collect the data. No effect-size calculations were done in this study. The author concludes by stating that: "Although the results of this experiment fail to demonstrate any superiority of reading over rote memorization (for either short-term or longer-term purposes), they nonetheless reveal that reading literature is at least as effective as rote memorization for the purpose of long-term vocabulary development" (Herman, 2003: 10).

Rasekh and Ranjbary (2003) seem to confuse the difference between a true experimental design and a quasi-experimental design. They state that: "This study had an intact group, pretest-posttest, experimental design" (Rasekh &

Ranjbary, 2003:8). This is not an experimental design, but a quasi-experimental design because the participants are used in intact groups and were not randomly assigned within the groups (cf. Hatch & Lazaraton, 1991). The results of this study indicate that the experimental group differed significantly from the control group on the vocabulary achievement test. "Thus, the explicit metacognitive strategy training seems to have contributed to the improvement of students' vocabulary learning" (Rasekh & Ranjbary, 2003: 11). This conclusion is based on the statistical significance of the results. No effect sizes were calculated in this study (cf. Table 9).

Although it is the submission policy of *Teaching English as a Second or Foreign Language* (cf. Appendix E) for authors to follow the APA guidelines when submitting articles for publication, all of studied reviewed ignored the requirement to conduct effect size analyses. The authors relied heavily only on statistical significance testing to either confirm or reject the null hypotheses formulated for the studies.

Table 9: Studies in Teaching English as a Second or Foreign Language (TESL-EJ)

Authors	Levine et al. (2000)
Research questions and/or hypotheses	• To what extent does the use of authentic tools, tasks and environment contribute to the development of critical reading skills and strategies in the computer networked EFL academic reading classroom in comparison to a conventional EFL academic reading class? • What is the nature of the EFL teacher's role in the computerized academic reading class? • What is the nature of the learners' behavior in the computerized EFL academic reading class and what is the effect of computer mediated instruction on their attitudes and motivation?
Participants	The subjects of the study were 58 Bar-Ilan University students of two groups of Advanced One level of English proficiency and two groups of Advanced Two level of English proficiency (note: students are placed in the two different levels of EFL courses on the basis of the National Psychometric Examination). Advanced One level students took a four-hour per week yearlong course and Advanced Two level students took a two-hour per week yearlong course.
Design	The research design consisted of two experimental and two control groups. The groups (experimental and control) were randomly selected: when registering for the course, the students were unaware which courses would eventually be taking part in the study. The decision regarding division into experimental and control groups was based on the availability of the networked computer classroom. The two experimental groups (16 Advanced One level students and 13 Advanced Two level students) studied exclusively in a networked class while the two control groups (16 Advanced One level students and 13 Advanced Two level students) were taught in a conventional classroom.
Data analysis techniques and statistics reported	Descriptive statistics (e.g., means, standard deviations) t-tests statistical significance

Intervention	The experimental classes studied according to the following study/lesson plan: During the first 10-15 minutes of class time, the students were asked to open the website and read a news item of their choice from one of the news links. The rest of the class time was devoted to individualized reading with students working independently on the topic and text of their choice. Each student could determine his/her pace of progress. At the beginning of the course, they were asked to skim each unit and its accompanying texts in order to choose a unit for focused reading. Students had the option of reading all the texts in the unit of their choice and then carry out the WebSearch assignment, or choose a unit, read all the texts and complete the corresponding worksheets before continuing to the WebSearch. The students were told that the purpose of the worksheets was to guide them through the reading process and to assist them in focusing on the ideas and information relevant to the topic of the unit.
Control	The two control groups studied in a conventional classroom. The students followed the same aim and scope of the course; they were taught by the same teacher and were provided with a hard copy of the same reading materials and the same worksheets that the students in the experimental groups received in electronic form. Daily news was downloaded from the Internet, printed and distributed to the students. The physical setting of the control groups was a conventional classroom. Each lesson began with 10-15 minutes of skimming through the headlines of daily downloaded news, followed by a short discussion of current events. After the initial reading and discussion, the students were given a text to read and were asked to do the assignments on the accompanying worksheet. The instructor, who also set the reading pace for the class, chose the text. Frontal instruction was provided for teaching and demonstrating as well as for overcoming reading problems, such as text structure, language, and questions on the worksheet.
Results	The findings of the study suggest that the computerized learning environment contributed to the development of EFL critical literacy skills to a greater extent than the conventional learning environment did. The advantages of the networked computer environment were evident mainly for EFL students at the higher level of language proficiency. The computer environment created a different teacher-student relationship and changed the nature of the EFL teacher's as well as the EFL student's role in the academic reading class. Differences in the application of close reading skills, global reading skills, and critical reading skills to the

	advantage of computerized classes were found both at the Advanced One and Advanced Two levels of English proficiency. The analysis of the independent T-test that was run to identify differences between the experimental (computerized) and control (conventional) groups revealed significant differences in the application of the following global reading skills: skimming of long texts in order to find the main ideas of the text : t(52) = 2.33*, p = .024 (p< .05) ; skimming in order to recognize the writer's purpose : t(52)= 2.73**, p=.002 (p<.01); skimming to identify the writer's conclusions : t(52)= 2.29*, p=.026 (p<.05).
Authors	Hermann (2003)
Research question and/or hypotheses	H_1: Subjects who participate in the reading of literature with a focus on meaning will acquire more vocabulary words than subjects who intentionally attempt to acquire the same words via memorization of paired associates. H_2: Subjects who acquire vocabulary incidentally through reading literature will exhibit higher retention rates than subjects who learn the same material through rote memorization.
Participants	Subjects in this study (N = 34) were enrolled in a freshman-level ESL composition course at a North American university. All subjects were fully matriculated students pursuing academic degrees. By virtue of their placement scores, they were judged to have writing abilities almost comparable to those of native English speakers enrolled at the freshman level. The language background of subjects in this study was quite heterogeneous, with a total of 13 native languages represented
Design	This study consists of a two-group, pre-test post-test quasi-experimental design. Since random sampling was not possible, four intact classes were chosen to serve as samples. All classes met three hours per week. To minimize the chances of inter-subject communication between groups, Tuesday/Thursday classes were chosen to function as the experimental group (n = 17), and Monday/Wednesday/Friday sections (n = 17) were selected to function as the comparison. Treatment in both groups was incorporated as part of the regular course curriculum for the respective sections; participation was therefore mandatory. However, due to low attendance on test days, several subjects were dropped from the study. Since repeated measures were performed, only data from subjects who took all three administrations of the vocabulary test were used. The n-sizes reflect the final status of both groups after all drops had occurred.
Data analysis	Descriptive statistics (e.g., means, standard deviations, median)

techniques and statistics reported	Spearman-rho analysis Statistical significance
Intervention	The experimental group was instructed to begin reading *Animal Farm*. Subjects in this group were instructed to focus on the novel's literary and rhetorical meaning and were informed they would be tested over major themes in the novel and their rhetorical development the following week. No instructions were given concerning vocabulary in the novel, and at no time were subjects told to expect a vocabulary test.
Control	After completing the pre-test, subjects were instructed to begin their designated assignments. For the comparison group, the assignment was to memorize a word list of paired-associates taken from *Animal Farm*. Subjects were unaware of the origin of the target words. Upon distribution of the list, subjects were informed that they would be tested on the words the following week. Subjects were not instructed to use any particular technique (e.g., the keyword method) but were merely told to commit the words to memory.
Results	Contrary to Hypothesis 1, the vocabulary gain from the pretest to the first post-test was vastly greater for the comparison group, $\chi^2 = 7.52$, $df = 1$, $p \leq .05$. In support of Hypothesis 2, however, the experimental group exhibited superior performance on the test for lexical retention, $\chi^2 = 17.18$, $df = 1$, $p \leq .05$. The comparison group actually experienced a 21% decline in mean performance between post-tests. Although the two groups performed quite differently with respect to each other, there was, interestingly enough, little difference between the groups' net vocabulary gains as measured from pretest to post-test 2: $\chi^2 = .47$, $df = 1$, $p \leq .49$.
Authors	Rasekh & Ranjbary (2003)
Research questions and/or hypotheses	Does metacognitive strategy instruction significantly increase the lexical knowledge of Iranian EFL students?

Participants	The participants of the study were 53 male and female Iranian EFL students taking part in an intensive course of English in Tehran Institute of Technology aged 19 to 25. The average age of the subjects was 21.86. They were studying English to enroll later in either English business classes or information technology classes. They had passed Headway elementary achievement test with at least 65 percent of the whole score. They were assigned into two classes and considered at pre-intermediate level of language proficiency. Twenty-nine of the subjects were female, and twenty-four were male. One of the classes was randomly selected as the control group and the other class as experimental group. To be sure of the homogeneity in the groups regarding lexical knowledge of the book that they were supposed to study during the experiment, a vocabulary achievement test (VAT) was developed and used. The reliability and validity of the vocabulary test was checked against Nelson language proficiency test administered to the subjects. The result of the vocabulary pre-test showed that there was no significant difference between the control and experimental groups in terms of lexical knowledge to be taught in the experiment period. The number of the students in the control was 26 and there were 27 subjects in the experimental group.
Design	This study had an intact group, pretest-posttest, experimental design. The subjects were already assigned in groups by the institution. Two classes were selected for this study and one was randomly assigned as experimental and the other as the control group.
Data analysis techniques and statistics reported	Descriptive statistics (e.g., means, standard deviations, standard error of the mean) Independent samples t-tests Statistical significance
Intervention	Both experimental and control groups enrolled in an English course which lasted for 10 weeks (four hours a day, three days a week). The textbook used for this course was Headway pre-intermediate. The authors have emphasized the role of lexical knowledge in learning the English language and have put some sections on vocabulary learning strategies in the book. One of the researchers taught both classes. Both groups received the usual training based on the procedures suggested in the Headway Teacher's book. The vocabulary

	strategies which were covered in the book were summarized and taught in the first session for both groups. The instruction and use of vocabulary learning strategies continued throughout the course for both groups of subjects. It is believed that metacognitive strategies are responsible for controlling other strategies and as a result they have their best effects if students are aware of other strategies that are available to them at the beginning of the course. Only the experimental group received explicit instruction on metacognitive strategies beginning from the second day of the course. The training was based on CALLA model of teaching learning strategy which includes five steps, namely preparation, presentation, practice, evaluation and expansion..
Control	The control group received training based on the procedures suggested in the Headway Teacher's book.
Results	The results of the vocabulary test in the two groups were compared using independent samples t-test statistical procedure, whose result showed that the mean scores of the experimental group (M = 29.29, SD = 3.84) was significantly (t (51) = 3.55 p< .05) different from the control group (M = 25.30, SD = 4.32). In other words, while there was not any significant difference between control and experimental group in terms of lexical knowledge at the beginning of the study, the experimental group surpassed the control group in terms of lexical knowledge at the end of the experiment.

6.3 Conclusion

A review of the literature on reading instruction research indicates that experimentation, when it is possible to randomly assign participants to differing forms of instruction, permits insights on instruction as a cause of achievement better than any other methodology (cf. Pressley, 2003). There is no substitute for the randomised experiment as a powerful window on cause-and-effect relationships. However, statements such as the following: "Intervention studies are also important because a recommendation for a change in teaching practice will be stronger if supported by evidence from a successful intervention study" (Nunes et al., 2003:289) should also be treated with caution. The reason is that many researchers overstate the claims they make based on their statistically significant results (cf. studies reviewed in this chapter). What can a researcher conclude when a new type of instruction produces positive achievement relative to existing instruction? At best, researchers can claim that the new form of instruction caused the achievement difference. According to Pressley (2003), the word *significant* has ambiguities associated with it when used to describe a research outcome. It can be used to state that the finding is statistically significant (e.g., there is less than a 5% or 1% or one tenth of 1% chance that the difference obtained is a chance difference). In this sense, a finding is considered to be highly significant when there is an exceptionally low probability that the difference observed was due to chance. Thus, when there is less than one tenth of 1% chance that a difference is due to chance, the finding is often referred to as highly significant even though such a statistical difference can be quite small in absolute or practical terms, so small as to affect performance hardly at all. Most often, when a small practical or absolute effect has high statistical significance, it is because there were a lot of participants in each of the instructional conditions. When that is the case, a difference of high statistical significance can be inconsequential in terms of educational or practical classroom significance. That is, the students receiving the new intervention do only a little bit better than the students receiving the conventional instruction.

According to several authors (e.g., Cohen, 1992; APA, 1994, 2001; Steyn, 2000; Pressley, 2003), it is, therefore, essential to know how big an effect the instruction under investigation had. Very few of the studies reviewed in this study gave heed to the APA (1994; 2001) call for authors to include effect size calculations for their results.

An additional aspect that emerged from the review of the studies is that many researchers are opting for a combination of quantitative and qualitative methodologies. This is an aspect that needs further attention as various authors (e.g., Purcell-Gates, 2000; Pressley, 2003) have stated that many educational issues and problems (e.g., reading instruction research) clearly require research that draws on multiple perspectives, approaches, and procedures. As researcher and educators we need methodologies that will allow us to probe for insights, for possible operative factors, for new information, before we can begin to think about limiting research studies to experimental, hypothesis-testing approaches.

CHAPTER 7
A FRAMEWORK FOR SELECTING, AND REPRESENTING STATISTICS IN QUANTITATIVE READING INSTRUCTION RESEARCH STUDIES

7.1 Introduction

Researchers as well as consumers of research need to do more than read the research question and then skip directly to the study findings and conclusions or implications. Although these elements of a research study can immediately inform the reader about what was investigated and whether there were significant findings, it is also important for researchers and consumers of research to be certain that the studies are methodologically appropriate for both the research questions guiding the study and the findings resulting from the study. Interpreting research would be a much easier process if it could always be safely assumed that studies were methodologically sound.

In recent years, the debate about the utilisation of research studies that have been used to inform educational practice, specifically reading instruction research, has been steadily growing (cf. Krashen, 2002; Pressley, 2003; 2006). There has been and continues to be much frustration among educators and parents alike; learner reading outcomes are not significantly improving. One possible explanation is that most reading instruction intervention studies are not informed by the same type of scientifically based research used in medical and other scientific fields (cf. Slavin, 2003). There is a growing call for concentrating future research investments in the field of education on experimental scientifically based research because other methods, specifically non-experimental studies, cannot identify effective practices and programmes. Proponents of scientifically based research in education argue that education is a science, like chemistry,

medicine, or psychology, and therefore is subject to the same rules of evidence as any other applied field (Shavelson & Towne, 2002).

The purpose of this chapter is to present a framework that seeks to provide assistance to reading educators in selecting and evaluating whether reading intervention studies are backed by rigorous evidence of effectiveness, especially with regard to the research results section, specifically the selection and presentation of the statistical results.

7.2 The proposed framework

Educators face many challenges in interpreting reading instruction research studies and their conclusions. The proliferation of research in this field is both encouraging and problematic because it cannot be assumed that a study's conclusions are valid simply because they are "published" (Dimsdale & Kutner, 2004). The sheer volume of research produced is daunting for any consumer, as is the variety of sources in which it can be found. Whereas the majority of research used to be published almost exclusively in scholarly journals, research is now published by think tanks, on websites, and by various organisations. Each source is subject to different sets of standards and varying levels of peer review. Unfortunately, bad research does get published and is not always given a "red flag" by the research community (McEwan & McEwan, 2003). In considering research, educators must be careful to understand how and for what purposes they can use the findings to inform their teaching practice. All research studies using an experimental and scientifically based research approach are not necessarily well designed and well implemented. Scientific jargon and explanations of advanced statistical procedures that accompany many research studies often add yet another obstacle to the process of understanding because educators do not always have training in advanced research methods to determine research quality (Dimsdale & Kutner, 2004).

Despite these challenges, with the appropriate knowledge educators can rely on research studies as effective and important sources of information to inform practice. Well-designed studies using an experimental and scientifically based approach are extremely useful tools for educators and administrators alike in improving reading instructional practices and programmes.

The framework presented in this chapter has three main parts (cf. Figure 7). The first part gives an outline of the aspects educators need to consider when critically reviewing a study's description of the intervention process as well as the randomization process followed with the participants. The second part focuses on the aspects that need to be considered when considering the data collection procedure used in the study. Part three focuses on the selection and reporting of the statistical techniques used in the results section of the study (cf. sections 4.2, 4.3 and 4.4).

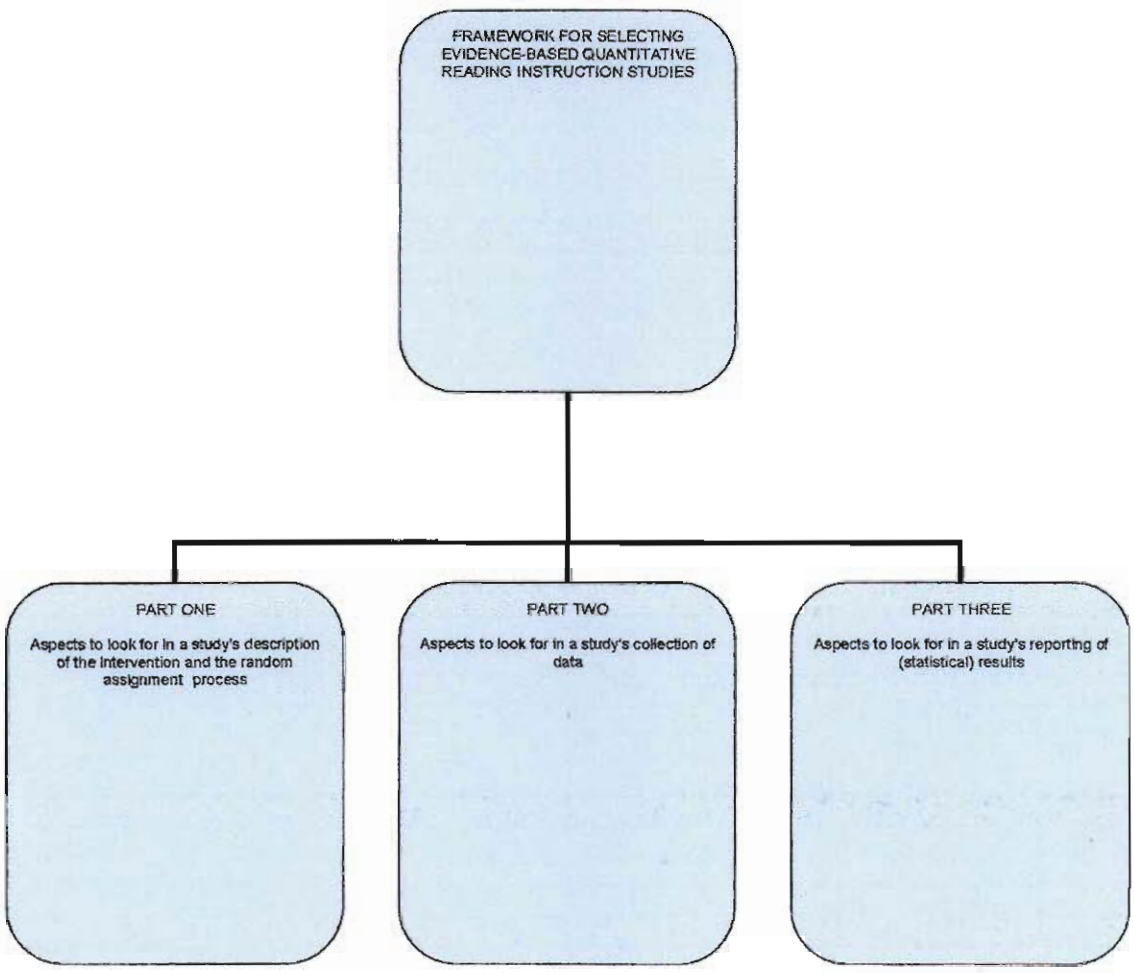


Figure 7: Framework for selecting evidence-based quantitative reading instruction studies

7.2.1 Part one: A description of the intervention and the random assignment process

Experimental studies are highly quantitative studies that attempt to empirically test a hypothesis by using "hard" data and statistical techniques (Ary et al., 2005). Results from experimental studies are primarily concerned with whether or not an intervention is effective. Experimental studies do not emphasize understanding how contextual factors and conditions influence outcomes. Data and findings produced by experimental studies provide the most definitive conclusions possible and thus are very appealing to policymakers, researchers, and stakeholders (Dimsdale & Kutner, 2004). Experimental or "true" designs are the gold standard for research because these studies attempt to establish causal relationships between two or more factors. A well-designed experimental study allows the researchers to answer a research question with a high degree of certainty because their conclusions are backed up by concrete data (Hatch & Lazaraton, 1991; Ary et al., 2005). The primary question an experimental study allows a researcher to answer is:

> Did the treatment or intervention have a significant effect on the treatment group? In the literacy field, for example, an experimental study can answer the question: Is phonics-based instruction an effective instructional approach for literacy learners?

Essential features that educators should focus on when evaluating the quality of a study during part one include (cf. section 2.3):

- Specifying the treatment – Although every experimental research study includes a section on the treatment used, the treatment is not always explained in an unambiguous way. Words such as "more" or "less" and "high" or "low" are not sufficiently descriptive. After reading about the treatment, the educator should understand exactly how the treatment group was treated differently from the control group. The treatment must be implemented consistently and the researchers should report measures

that demonstrate treatment fidelity (Lauer, 2004). The study should clearly describe (i) the intervention, including who administered it, who received it, and what it cost, (ii) how the intervention differed from what the control group received, and (iii) how the intervention is supposed to affect outcomes (Barnette, 2006) (cf. section 2.3).

For example, a randomized controlled trial of a one-on-one Direct Instruction programme for grade one to grade three readers should discuss such items as:

- who conducted the training (e.g., certified teachers, undergraduate volunteers, etc.);
- what training they received in how to tutor;
- what curriculum they used to tutor, and other key features of the training sessions (e.g., daily 20-minute sessions over a period of six-months);
- the age, reading achievement levels, and other relevant characteristics of the tutored students and controls;
- the cost of the training intervention per student;
- the reading instruction received by the students in the control group (e.g., the school's pre-existing reading programme); and
- the logic by which training is supposed to improve reading outcomes (cf. American Federation of Teachers, 1999).

- Control group – All experimental studies are performed by comparing two groups: a treatment group that is exposed to some kind of intervention and a control group that does not receive the intervention. The control group provides the groundwork for a comparison to be made and is perhaps the most essential feature of experimental research. It gives researchers an insight into how the treatment group would have theoretically behaved if group members had never received the treatment, and therefore it allows researchers to judge the effectiveness of the treatment (Hatch &

Lazaraton, 1991; Ary et al., 2005). A "significant" improvement within only one group can be the result of time or other factors than the treatment, therefore, the comparison with a control group is essential to verify the effect of the treatment.

- Random Assignment – The use of comparison groups alone does not qualify a study as a true experiment. The participants in the study must also be randomly assigned to either the experimental group or the control group to ensure that there are no pre-existing differences between the groups (Ary et al., 2005). Randomly assigning participants to one of the two groups allows researchers to say with confidence that any differences shown between the groups after the intervention was a result of the intervention alone. If instead of being randomly assigned to the treatment and control groups, participants volunteer or are selected through a "convenience" sample, the experimental nature of the study is compromised (Dimsdale & Kutner, 2004; Ary et al., 2005). Educators should also be alert as to any indication that the random assignment process may have been compromised. For example, did any individuals randomly assigned to the control group subsequently cross over to the intervention group? Or did individuals unhappy with their prospective assignment to either the intervention or control group have an opportunity to delay their entry into the study until another opportunity arose for assignment to the preferred group? Such self-selection of individuals into their preferred groups undermines the random assignment process, and may well lead to inaccurate estimates of the intervention's effects. Ideally, a study should describe the method of random assignment it used (e.g., use of random numbers), and what steps were taken to prevent undermining (e.g., asking an objective third party to administer the random assignment process) (Pattern, 2002; Ary et al., 2005; What Works Clearinghouse, 2006b).

The study should provide data showing that there were no systematic differences between the intervention and control groups before the intervention. As discussed above, the random assignment process ensures, to a high degree of confidence, that there are no systematic differences between the characteristics of the intervention and control groups prior to the intervention (Ary et al., 2005). The degree of confidence increases with the group sizes. For small groups differences are bound to occur. It should be noted that there are other random designs than the two group design (cf. Hatch & Lazaraton, 1991; Ary et al., 2005).

- Non-contamination of participants – One of the greatest threats to the internal validity of an experimental study is contamination of participants. This can occur when the participants of the experimental and control groups communicate or when the instruction in the control group adopts key aspects of the treatment instruction (Pattern, 2002).

7.2.2 Part two: Data collection procedure

Experimental studies in reading are used to determine the impact of a reading intervention or programme. These impacts are typically learner outcomes: how learners' knowledge, skills, or performance have changed or improved as a result of the intervention. Standardized assessments are therefore the most commonly used data collection instrument in experimental studies (Marzano, Pickering, & Pollock, 2001). A successful experimental study should use an assessment instrument that directly measures the skills being taught by the intervention. Reliability refers to the consistency of the results of an assessment and is a necessary but not sufficient condition for validity. If an assessment used to measure reading comprehension is given to the same student three times without any instruction to the student, you would expect the student's test score to be about the same each time. If it was not, the instrument would not be providing reliable measures (Hatch & Lazaraton, 1991; Marzano et al., 2001; Ary et al., 2005). If possible the Cronbach alpha coefficients should be determined for

the data to establish internal consistency of items in a dimension/field/factor. Also, construct validity can be checked by means of a factor analysis. At a minimum, assessments must be administered to both experimental and control group participants prior to the beginning of the intervention (the pre-test) and at the conclusion of the intervention (the post-test). Assessments may also be administered to both the experimental and control groups periodically during the intervention period, as well as at a predetermined period of time after the intervention has been completed if the study's objective is to measure long-term programme impacts (Dimsdale & Kutner, 2004) (cf. section 2.3).

Wherever possible, a study should use objective, "real-world" measures of the outcomes that the intervention is designed to affect. If outcomes are measured through interviews or observation, the interviewers/observers preferably should be kept unaware of who is in the intervention and control groups. Such "blinding" of the interviewers/observers, where possible, helps protect against the possibility that any bias they may have (e.g., as proponents of the intervention) could influence their outcome measurements. When study participants are asked to "self-report" outcomes, their reports should, if possible, be corroborated by independent and/or objective measures. Thus, studies that use such self-reported outcomes should, if possible, corroborate them with other measures (e.g., third-party observations) (Pattern, 2002; Ary et al., 2005).

During experimental research studies it is also important to measure and document that the intervention is being delivered as designed. This is especially important when an experimental study involves multiple sites. Site visits by researchers to observe instruction and instructional logs completed by the instructors of both experimental and control groups are useful data-collection instruments to measure the fidelity of the intervention and to ensure that instruction received by the control group participants is different from the instruction received by experimental group participants throughout the period of study (Dimsdale & Kutner, 2004).

The percent of study participants that the study has lost track of when collecting data should be small, and should not differ between the intervention and control groups. A general guideline is that the study should lose track of fewer than 25 % of the individuals originally randomized; the fewer lost, the better. This is sometimes referred to as the requirement for "low attrition." (Studies that choose to follow only a representative sub-sample of the randomized individuals should lose track of less than 25% of the sub-sample). Furthermore, the percentage of subjects lost track of should be approximately the same for the intervention and the control groups. This is because differential losses between the two groups can create systematic differences between the two groups, and thereby lead to inaccurate estimates of the intervention's effect. This is sometimes referred to as the requirement for "no differential attrition" (What Works Clearinghouse, 2006b).

The study should preferably obtain data on long-term outcomes of the intervention, so that you can judge whether the intervention's effects were sustained over time. This is important because the effect of many interventions diminishes substantially within 2-3 years after the intervention ends. This has been demonstrated in randomized controlled trials in areas such as early reading. In most cases, it is the longer-term effect, rather than the immediate effect, that is of greatest practical and policy significance (Baron, 2004).

7.2.3 Part three: Selecting and reporting results (specifically statistics) in reading instruction research

If the study makes a claim that the intervention is effective, it should report (i) the size of the effect (cf. section 4.2), and (ii) statistical tests showing the effect is unlikely to be the result of chance (cf. sections 3.3, 4.3 and 4.4). Specifically, the study should report the size of the difference in outcomes between the intervention and control groups (What Works Clearinghouse, 2006a, 2006b). Tables 10 to 12 contain a menu that researcher and educators can use when

selecting and reporting statistical results. The following essential features should be noted and considered when consulting the tables:

Independent vs dependent variables

In an experimental design, the independent variable (also called the explanatory variable) is the variable which is manipulated or selected by the experimenter to determine its relationship to an observed phenomenon (the dependent variable). In other words, the experiment will attempt to find evidence that the values of the independent variable determine the values of the dependent variable (which is what is being measured) (Hatch & Lazaraton, 1991; Ary et al., 2005).

More generally, the independent variable is the thing that someone actively controls/changes; while the dependent variable is the thing that changes as a result. In other words, the independent variable is the "presumed cause", while dependent variable is the "presumed effect" of the independent variable.

Scales of measurement

Statistical information, including numbers and sets of numbers, has specific qualities that are of interest to researchers. These qualities, including magnitude, equal intervals, and absolute zero, determine what scale of measurement is being used and therefore what statistical procedures are best. Magnitude refers to the ability to know if one score is greater than, equal to, or less than another score. Equal intervals means that the possible scores are each an equal distance from each other. And finally, absolute zero refers to a point where none of the scale exists or where a score of zero can be assigned (Ary et al., 2005).

When we combine these three scale qualities, we can determine that there are four scales of measurement. The lowest level is the nominal scale, which represents only names and therefore has none of the three qualities. A list of students in alphabetical order, a list of favourite cartoon characters, or the names on an organizational chart would all be classified as nominal data. The second

level, called ordinal data, has magnitude only, and can be looked at as any set of data that can be placed in order from greatest to lowest but where there is no absolute zero and no equal intervals. Examples of this type of scale would include Likert Scales and the Thurstone Technique (Brown, 1988; Hatch & Lazaraton, 1991; Ary et al., 2005).

The third type of scale is called an interval scale, and possesses both magnitude and equal intervals, but no absolute zero. Temperature is a classic example of an interval scale because we know that each degree is the same distance apart and we can easily tell if one temperature is greater than, equal to, or less than another. Temperature, however, has no absolute zero because there is (theoretically) no point where temperature does not exist (Brown, 1988; Hatch & Lazaraton, 1991; Ary et al., 2005).

Finally, the fourth and highest scale of measurement is called a ratio scale. A ratio scale contains all three qualities and is often the scale that statisticians prefer because the data can be more easily analyzed. Age, height, weight, and scores on a 100-point test would all be examples of ratio scales. If you are 20 years old, you not only know that you are older than someone who is 15 years old (magnitude) but you also know that you are five years older (equal intervals). Ratios also make sense: the 20 year old is one and a third times as old as the 15 year old. With a ratio scale, we also have a point where none of the scale exists; when a person is born his or her age is zero (Brown, 1988; Hatch & Lazaraton, 1991; Ary et al., 2005).

Independent or repeated levels

Another important aspect to consider when thinking about variables is levels. A nominal variable, by definition, can include a number of categories. These categories are also called levels (Brown, 1992b: 634). For example, for a variable like gender, there are two levels, female and male. For a variable like language proficiency, there might be three levels, elementary, intermediate and advanced. This definition of levels is important in thinking about statistical studies because

of the concept of independence. If the levels of a variable are made up of two or more different groups of people (i.e. the levels are mutually exclusive), they are viewed as independent. For example, in a study comparing the means of an experimental group and a control group, the groups can be independent only if they were created using different groups of subjects. Many statistical analyses can only be applied if the groups being compared are independent.

However, there might be studies in which it is desirable or necessary to make repeated observations on the same group of subjects. For example, a researcher might want to compare the means of a single group of students before and after some type of reading instruction in what is called a pre-test post-test study. Such investigations are called repeated measures studies, and the groups lack independence because they are the same subjects. When independence cannot be assumed, different choices of statistical analysis techniques must be made (Brown, 1988, 1992a, 1992b; Hatch & Lazaraton, 1991; Ary et al., 2005).

Other conditions/aspects

In statistics, a covariate is a variable that is possibly predictive of the outcome under study. A covariate may be of direct interest or be a confounding variable or effect modifier. It is also sometimes referred to as a covariable; an independent variable, or predictor, in a regression equation. Also, a secondary variable that can affect the relationship between the dependent variable and independent variables of primary interest in a regression equation (Brown, 1988, 1992a, 1992b; Hatch & Lazaraton, 1991; Ary et al., 2005).

In order to obtain such a finding of statistically significant effects, a study usually needs to have a relatively large sample size. A rough rule of thumb is that a sample size of at least 300 students (150 in the intervention group and 150 in the control group) is need to obtain a finding of statistical significance for an intervention that is modestly effective. If schools or classrooms, rather than

individual students, are randomized, a minimum sample size of 50 to 60 schools or classrooms (25-30 in the intervention group and 25-30 in the control group) is needed to obtain such a finding. (This rule of thumb assumes that the researchers choose a sample of individuals or schools/classrooms that do not differ widely in initial achievement levels.). If an intervention is highly effective, smaller sample sizes than this may be able to generate a finding of statistical significance (Hatch & Lazaraton, 1991; Ary et al., 2005). If the study seeks to examine the intervention's effect on particular subgroups within the overall sample (e.g., Setswana-speaking students), larger sample sizes than those above may be needed to generate a finding of statistical significance for the subgroups (cf. sections 3.2.3.2 and 4.3.3).

In general, larger sample sizes are better than smaller sample sizes, because they provide greater confidence that any difference in outcomes between the intervention and control groups is due to the intervention rather than chance. If the study randomizes groups (e.g., schools) rather than individuals, the sample size that the study uses in tests for statistical significance should be the number of groups rather than the number of individuals in those groups. Then the means of the randomized groups will be treated as the data. Occasionally, a study will erroneously use the number of individuals as its sample size, and thus generate false findings of statistical significance (Ary et al., 2005).

For example, if a study randomly assigns two schools to an intervention group and two schools to a control group, the sample size that the study should use in tests for statistical significance is just four, regardless of how many hundreds of students are in the schools. (And it is very unlikely that such a small study could obtain a finding of statistical significance). The study should preferably report the size of the intervention's effects in easily understandable, real-world terms (e.g., an improvement in reading skill by two grade levels). It is important for a study to report the size of the intervention's effects in this way, in addition to whether the effects are statistically significant, so that the reader (e.g., educator) can judge

their educational importance. For example, it is possible that a study with a large sample size could show effects that are statistically significant but so small that they have little practical or policy significance (e.g., a 2 percent increase on a standardized reading test). Unfortunately, some studies report only whether the intervention's effect is statistically significant, and not their magnitude.

Selecting statistical techniques and effect-size measures

This decision is a function of the research question asked and the nature and the level of measure of the variables involved in the study. That is, educators should begin with the research question, identify the dependent and independent variables involved, identify the level of measurement of every variable, and then to Tables 10-12 that will point them to the appropriate technique.

Assumptions underlying the use of statistical tests

Brown (1992b: 639) states that assumptions are preconditions that are necessary for accurate application of a particular statistical test. In some cases, these assumptions are not optional; they must be met for the statistical test to be meaningful.

For all cases where the dependent variable is interval and the samples are small, normality has to be assumed, otherwise the nonparametric methods (as with ordinal DVs) have to be used. Normality is a probability distribution that has a symmetrical bell-shaped curve. The data are evenly distributed around the mean trait value. Non-normally distributed variables (highly skewed or kurtotic variables, or variables with substantial outliers) can distort relationships and significance tests. There are several pieces of information that are useful to the researcher in testing this assumption: visual inspection of data plots, skew, kurtosis, and P-P plots give researchers information about normality, and Kolmogorov-Smirnov tests provide inferential statistics on normality. Outliers can be identified either through visual inspection of histograms or frequency distributions, or by converting data to z-scores.

An assumption, when dealing with ANOVAs, is that of equal standard deviations for the groups (called homoscedasticity). Homogeneity of variances (or the attribute of not being statistically different) between the different experimental groups (or treatments) being compared. This can be determined using Bartlett's Test for Homogeneity of Variances if there are more than two treatment groups. This is one of the three assumptions underlying statistical tests along with random sampling and a normal distribution.

Table 10: Selection menu for group comparisons

Number and scale of Dependent Variable (DV)	Number and scale of Independent Variables (IV)	Independent or Repeated measures	Number of levels of IV	Other conditions/aspects	Appropriate statistic	Strength of Association/Effect size measure
1 Interval DV	1 Nominal IV	Independent	2 levels	Large sample Any sample	Z statistic T test	Cohen's effect size d Omega^2
			2 or more levels	No covariate Covariate	One-way ANOVA One-way ANCOVA	
		Repeated	2 levels	No covariate	Matched-pairs t-test Repeated measures ANOVA	Eta^2
			2 or more levels	Covariate	Repeated measures ANCOVA (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	(Cohen, 1988; Hatch & Lazaraton, 1991; Wikipedia, 2006).
I Interval DV	2 or more (n) Nominal IV	Independent		No covariate	n-way ANOVA	Eta^2
				Covariate	n-way ANCOVA	
		Repeated		No covariate	n-way repeated measures ANOVA	
				Covariate	n-way repeated measures ANCOVA (Hatch & Lazaraton, 1991; Statsoft, 2005;	

					Wikipedia, 2006)	
2 or more Interval DV's	1 Nominal IV	Independent	2 levels	No covariate	Hotelling's T^2	Eta^2
			2 or more levels	No covariate	Multivariate ANOVA	
				Covariate	Multivariate ANCOVA	
	2 or more (n) Nominal IV	Independent		No covariate	Multivariate n-way ANOVA	
				Covariate	Multivariate n-way ANCOVA	
					(Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	
1 Ordinal DV	1 Nominal IV	Independent	2 levels		Wilcoxon 2-sample test	Eta^2
			3 or more levels		Kruskal-Wallis test	
		Repeated	2 levels		Wilcoxon signed rank test	
			3 or more levels		Friedman one-way ANOVA	
					(Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	

Table 11: Selection menu for correlation/prediction studies

Number and scale of Dependent Variable (DV)	Number and scale of Independent Variables (IV)	Number of levels of IV	Appropriate statistic	Strength of Association/Effect size measure
1 Interval DV	1 Interval IV 2 or more Interval IV		Simple linear regression and Pearson r Multiple linear regression and multiple R (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	Cohen's effect size r^2 or R^2 (Cohen, 1988; Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)
2 or more Interval DV	2 or more Interval IV		Canonical correlation (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	
1 Ordinal DV	1 Ordinal IV	1 level 2 or more levels	Spearman rho or Kendall tau Kendall W (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	
1 Nominal DV on 2 levels	1 Interval IV	True dichotomy Artificial dichotomy	Point-biserial correlation Biserial correlation (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	r^2 r^2 (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)
1 Nominal DV on 2 levels	1 Nominal IV	True dichotomy Artificial dichotomy	Phi coefficient Tetrachoric correlation (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	Phi coefficient (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)

Table 12: Selection menu for frequency statistics

Number and scale of Dependent Variable (DV)	Number and scale of Independent Variables (IV)	Independent or Repeated measures	Appropriate statistic	Strength of Association/Effect size measure
1 Nominal DV on 2 levels	1 Nominal IV	Independent Repeated	Chi-square, Fisher's exact test McNemar test (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	Phi Cramer's V (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)
1 Nominal DV	2 or more Nominal IV	Independent	n-way chi-square, multiway frequency analysis (Hatch & Lazaraton, 1991; Statsoft, 2005; Wikipedia, 2006)	Cohen's w (Cohen, 1988)

7.3 Conclusion

With attention increasingly being paid to scientifically based reading research, educators are already facing the results from new research studies, and more and more of the studies are based on experimental or quasi-experimental research designs. The key is for educators to take a critical and objective look at research findings and determine whether the study methodology is appropriate, both in design and implementation, for the research questions being asked and the study findings being presented.

CHAPTER 8

CONCLUSION AND RECOMMENDATIONS FOR FUTURE RESEARCH

8.1 Introduction

It is imperative now and in the future that reading professionals have access to complete and thorough syntheses of findings from a broad base of research on reading-related topics from multiple types of studies. Just as important is the challenge to help educators understand their roles in scientific research. The challenge is to engage educators in ongoing study of the changing knowledge base produced by advances in scientific reading research so that they can make informed decisions and contribute to locally developed, scientifically based programmes of instruction and assessment in reading.

Educators cannot be mere consumers of the knowledge emerging from the interrelated and continuously advancing systems of scientific reading research. They must position themselves as active participants in the research community. They must understand that their own honing of "proven" instructional strategies through reflection and systematic monitoring of students' progress is a critical component of a scientifically based reading programme. Engaging in scientific reading research is about extending our knowledge through inquiry by carefully observing, generating, and testing hypotheses; collecting data; and drawing evidence-based conclusions that are shared with other reading educators and researchers.

The purpose of this chapter is to summarise the key aspects addressed in this study.

8.2 Results

The following research questions were formulated in chapter one of this study:

- What is the state of affairs with regard to statistical significance testing in reading instruction research, with specific reference to post-1999 literature?
- What are the criticisms as well as the defences that have been offered for statistical significance testing?
- What are the alternatives or supplements to statistical significance testing that may be relevant for reading instruction research?

With regard to the first research question, the results indicated the following: A review of six readily accessible (online) journals publishing research on reading instruction indicated that researchers/authors rely very heavily on statistical significance testing and very seldom, if ever, report effect size/effect magnitude or confidence interval measures when documenting their results (cf. Chapter 6). This happens despite the editorial policies of some of these journals requiring adherence to the APA guidelines for reporting quantitative research results (i.e., the inclusion of effect size measures) (cf. Appendices A-F).

Research questions two and three were discussed in detail in chapters 3 and 4. A review of the literature indicates that:
Null hypothesis significance testing has been and is a controversial method of extracting information from experimental data and of guiding the formation of scientific conclusions. Meehl (1997:421) states that:

> Competent scholars persist in strong disagreement, ranging from some who think H_0-testing is pretty much all right as practices, to others who think it is never appropriate. Most critics fall somewhere between these extremes, and they differ among themselves as to their main *reasons* for complaint.

Each of the suggested alternatives or complements to null hypothesis significance testing, namely effect sizes, confidence intervals and power

analysis has its proponents. In trying to assess the merits of null hypothesis significance testing and other approaches to the evaluation of hypotheses, it is well to bear in mind that nothing of importance in reading instruction research has ever been decided on the basis of the outcome of a single statistical significance test. Reading instruction knowledge is acquired as a consequence of the cumulative effect of many experiments and non-experimental observations as well. It is the preponderance of evidence gathered from many sources and over an extended period of time that determines the degree of credibility that is given to hypotheses, models and theories.

8.3 Central theoretical statement

The following central theoretical statement was formulated for this study:

> Statistical significance tests should be supplemented with accurate reports of effect size, power analyses and confidence intervals in reading research studies. In addition, quantitative studies, utilising statistics as stated in the previous sentence, should be supplemented with qualitative studies in order to obtain a more comprehensive picture of reading instruction research.

Research indicates that no single study ever establishes a programme or practice as effective; moreover it is the convergence of evidence from a variety of study designs that is ultimately scientifically convincing. When evaluating studies any claims of evidence, educators must not determine whether the study is quantitative or qualitative in nature, but rather if the study meets the standards of scientific research. That is, does it involve "rigorous and systematic empirical inquiry that is data-based" (Bogdan & Biklen, 1992:43).

Patton (1990:39) advocates a "paradigm of choices" that seeks "methodological appropriateness as the primary criterion for judging methodological quality". Furthermore, some researchers believe that qualitative and quantitative research can be effectively combined in the same research project (Patton, 1990; Strauss & Corbin, 1990). For

example, Russek and Weinberg (1993) claim that by using both quantitative and qualitative data, their study of technology-based materials for the elementary classroom gave insights that neither type of analysis could provide alone.

8.4 The proposed framework

The proposed framework presented in this study consists of three main parts:

In part one the key aspects to look for in the study's description of the intervention and the random assignment process include:

The study should clearly describe the intervention, including: (i) who administered it, who received it, and what it cost; (ii) how the intervention differed from what the control group received; and (iii) the logic of how the intervention is supposed to affect outcomes. Educators should be alert to any indication that the random assignment process may have been compromised. The study should provide data showing that there are no systematic differences between the intervention and control groups prior to the intervention.

In part two the key aspects to look for in the study's collection of data include:

The study should use outcome measures that are "valid" (i.e., that accurately measure the true outcomes that the intervention is designed to affect). The percent of study participants that the study has lost track of when collecting outcome data should be small, and should not differ between the intervention and control groups. The study should collect and report outcome data even for those members of the intervention group who do not participate in or complete the intervention. The study should preferably obtain data on long-term outcomes of the intervention, so that you can judge whether the intervention's effects were sustained over time.

In part three the key aspects to look for in the study's reporting of results include:

If the study makes a claim that the intervention is effective, it should report (i) the size of the effect, and (ii) statistical tests showing the effect is unlikely to be the result of chance. In part three the focus is on the presentation of selection menu's that can be used by educators to select appropriate statistical tests as well as effect size measures.

The effect size reporting creates a literature in which subsequent researchers can more easily formulate more specific study expectations by integrating the effects reported in related prior studies. In addition, interpreting the effect sizes in a given study facilitates the evaluation of how a study's results fits into existing literature, the explicit assessment of how similar or dissimilar results across related studies, and potentially judgment regarding what study features contributed to similarities or differences in effects. The fifth edition of the APA Publication Manual now incorporates as a requirement, "**Always** provide some effect size estimate when reporting a p value" (Wilkinson & APA Task Force on Statistical Inference, 1999: 599) (my emphasis).

8.5 Conclusion

The quest to find the "best programmes" for teaching reading has a long and quite unsuccessful history (International Reading Association, 2002). Despite many claims of programme excellence, literacy scholars (e.g., Allington, 2001; Stahl et al., 1998) argue that careful examination of such studies reveals the use of either flawed designs or selective reporting of the available data.

The challenge that confronts educators and administrators is the need to view the evidence that they read through the lens of their particular school and classroom settings. They must determine if the instructional strategies and routines that are central to the materials under review are a good match for the particular learners they teach.

Pearson (quoted in The National Right to Read Foundation, 2003) commented on the need for teacher educators to prepare teachers who are knowledgeable about evidence-based reading instruction. "Somehow we must generate the political will to insist that every teacher receive quality training on all aspects of research-based pedagogy".

Each and every teacher education programme in South Africa should ensure that its students know how to read reports of the very latest research so that they can change their classroom practices to match new evidence.

8.6 Recommendations for future research

If the goal of scientific inquiry in the reading instruction research field is to determine if the results of a test have any practical importance, it is recommended that all quantitative research utilizing statistical significance testing report an effect magnitude measure to highlight the distinction between statistical and practical significance.

The adoption and enforcement of more strict editorial policies regarding the reporting of the results of statistical significance testing and effect magnitude measures will perhaps eventually move the field toward improved practice. Editorial policies in the journals focusing on reading instruction should (a) require authors to index results of statistical significance tests to sample size, (b) require effect magnitude measures with statistical significance tests, and (c) encourage the use and reporting of confidence intervals.

Miller (1998) suggested that if statistics are the tools of the researcher, we, as researchers then need to know our tools: "Tractor mechanics, artists, and masons have their tools and they must know how to use them. We, likewise, need to know how to use ours. You are challenged to get 'checked-out' again on your tools; that is, devote some of your personal in-service or professional development time to renewing, maintaining, and improving your skills" (Miller, 1998:1).

BIBLIOGRAPHY

ABELSON, R.P. 1997. A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). *In* Harlow, L.L., Mulaik, S.A. & Steiger, J.H. (Eds.). (What if there were no significance tests?) Mahwah, NJ: Erlbaum.

ALBANESE, M.A. & MITCHELL, S. 1993. Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic medicine*, 68:52-81.

ALLINGTON, R.L. 2001. What really matters for struggling readers. Designing research-based programs. New York: Longman.

AL-SEGAHYER, K. 2001. The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language learning & technology*, 5(1):202-232.

AMERICAN COLLEGE OF PHYSICIANS.. 2001. Primer on 95% confidence intervals. *Effective clinical practice*, 4(5):229-231.

AMERICAN FEDERATION OF TEACHERS. 1999. Building on the best, Learning from what works. Five promising remedial reading intervention programs. Washington, D.C.: American Federation of Teachers.

AMERICAN PSYCHOLOGICAL ASSOCIATION. 1994. Publication manual of the American Psychological Association 4th ed. Washington, DC: Author.

AMERICAN PSYHOLOGICAL ASSOCIATION. 2001. Publication manual of the American Psychological Association. 5th ed. Washington, D.C.: Author.

ANDERSON, D.R., BURNHAM, K.P. & THOMPSON, W.L. 1999. Null hypothesis testing in ecological studies: Problems, prevalence, and an alternative. Paper presented at the annual meeting of the Southwest Educational Research Association. San Antonio, Texas.

ANDERSON, D.R., BURNHAM, K.P. & THOMPSON, W.L. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of wildlife management*, 640:912-923.

ARON, A. & ARON, E.N. 2003. Statistics for psychology. 3rd ed. New Jersey, Upper Saddle River: Prentice-Hall.

ARY, D. JACOBS, L.C. & RAZAVIEH, A. 2005. Introduction to research in education. 7th ed. California: Wadsworth/Thomson Learning.

ASRAF, R.M. & BREWER, J.K. 2004. Conducting tests of hypotheses: The need for an adequate sample size. *The Australian Educational Researcher,* 31(1):79-94.

BAKAN, D. 1966. The test of significance in psychological research. *Psychological bulletin,* 66:423-437.

BARNETTE, J.J. 2006. Effect size and measures of association. Paper presented at the summer evaluation institute, Alabama. USA.

BARON, J. 2004. Identifying and Implementing education practices supported by rigorous evidence: A User Friendly Guide. *The journal for vocational special needs education*, 26(2):40-54.

BAUGH, F. 2002. Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and psychological measurement*, 62(2):254-263.

BECKER, G. 1991. Alternative methods of reporting research results. *American psychologist*, 46:654-655.

BELL, T. 2001. Extensive reading: speed and comprehension. *The reading matrix*, 1(1):1-13.

BERKSON, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, 37:325-335.

BOGDAN, R.C. & BIKLEN, S.K. 1992. Qualitative research for education: An introduction to theory and methods. Boston: Allyn & Bacon.

BORING, E.G. 1919. Mathematical vs. scientific importance. *Psychological bulletin*, 16(10):335-338.

BOSTON, C. 2003-2004. Effect size and meta-analysis http://www.ericdigests.org/2003-4/meta-analysis.html. Date of access: 3 Oct. 2005.

BRANTMEIER, C. 2004. Statistical procedures for research on L2 reading comprehension: An examination of ANOVA and regression models. *Reading in a foreign language*, 16(2):51-69.

BROWN, J.D. & RODGERS, T. 2002. Doing applied linguistics research. Oxford: Oxford University Press.

BROWN, J.D. 1988. Understanding research in second language learning. New York: Cambridge University Press.

BROWN, J.D. 1991. Statistics as a foreign language Part 1: What to look for in reading statistical language studies. *TESOL quarterly*, 25(4):569-586.

BROWN, J.D. 1992a. What is research? *TESOL matters*, 2:10.

BROWN, J.D. 1992b. Statistics as a foreign language. Part 2: More things to look for in reading statistical language studies. *TESOL quarterly*, 26:629-664.

BROWN, J.D. 2004. Resources on quantitative/statistical research for applied linguistics. *Second Language Research*, 20(4):374-393.

CARVER, R. 1993. The case against statistical significance testing, revisited. *Journal of experimental education*, 61: 287-292.

CARVER, R.P. 1978. The case against statistical significance testing. *Harvard educational review*, 48:378-399.

CASTIGLIONI-SPALTEN, M.L. & EHRI, L.C. 2003. Phonemic awareness instruction: Contribution of articulatory segmentation to novice beginners' reading and spelling. *Scientific studies of reading*, 7(1):25-52.

CHAUDRON, C. 1988. *Second language classrooms: Research on teaching and learning*. Cambridge: Cambridge University Press.

CHRISTENSEN, C.A. & BOWEY, J.A. 2005. The Efficacy of orthographic rime, grapheme-phoneme correspondence, and implicit phonics Approaches to teaching decoding skills. University of Queensland.

Coalition for evidence-based policy. 2003. *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide.* Washington, DC: U.S. Department of Education. http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf. Date of access: 13 May, 2004.

COE, Robert. Sept. 2002. It's the Effect Size, Stupid. What effect size is and why it is important. Paper presented at the annual conference of the British Educational Research Association, University of Exeter, England.

COHEN, J. 1988. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erbaum.

COHEN, J. 1990. Things I have learned (so far). *American psychologist*, 45:1304-1312.

COHEN, J. 1992. A power primer. *Psychological bulletin*, 112:155-159.

COHEN, J. 1994. The earth is round ($p < .05$). *American psychologist*, 49:997-1003.

COMPREHENSIVE SCHOOL REFORM PROGRAM OFFICE. 2002. *Scientifically based research and the Comprehensive School Reform (CSR) Program*. Washington, DC: U.S. Department of Education. http://www.ed.gov/programs/compreform/guidance/appendc.pdf. Date of access: 13 May, 2004.

CORTINA, J.M. & DUNLAP, W.P. 1997. Logic and purpose of significance testing. *Psychological methods,* 2:161-172.

CUMMING, G. & FINCH, S. 2001. A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and psychological measurement*, 61:530-572.

DANIEL, L.G. 1997. Statistical significance testing in "Educational and psychological measurement" and other journals. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago: Illinois.

DANIEL, L.G. 1998. Statistical Significance testing: A historical overview of misuse and miss-interpretation with implication for the editorial policies of educational journals. *Research in the schools*, 5(2):23-32.

DAVIES, H.T.O. 2001. What are confidence intervals? Kent: Hayward Medical Communications.

DENIS, D.J. 2003. Alternatives to null hypothesis significance testing. *Theory & science*, 4(1):1-23.

DENTON, C.A., VAUGHN, S. & FLETCHER, J.M. 2003. Bringing research-based practice in reading intervention to scale. *Learning disabilities research & practice,* 18(3):201-211.

DIMSDALE, T. & KUTNER, M. 2004. A Quick Look at the basics of research methodologies and design. American Institutes for research.

DIMSDALE, T. & KUTNER, M. 2004. Becoming an Educated Consumer of Research: A Quick Look at the Basics of Research Methodologies and design. American Institutes for Research. www.air.org. Date of access: 4 January, 2005.

DIXON, P. 1998. Why scientists value p values. *Psychonomic bulletin and review*, 5:390-396.

DOCHY, F., SEGERS, M. & BUEHL, M.M. I999. The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of educational research*, 69(2):145-186.

DOWSON, V. 2000. "Time of day effects in school-children's immediate and delayed recall of meaningful material". *TERSE Report* . http://www.cem.dur.ac.uk/ebeuk/research/terse/library.htm. Date of access: 9 November, 2004.

DREYER, C. & NEL, C. 2003. Teaching reading strategies and reading comprehension within a technology-enhanced learning environment. *System*, 31:349-365.

DREYER, C. 1998. Improving students' reading comprehension by means of strategy instruction. *Journal for language teaching,* 32(1):18-29.

FAN, X. 2001. Statistical significance and effect size in education research: Two sides of a coin. *The journal of educational research,* 94(5):275-282.

FARSTRUP, A.E. 2003. The importance of evidence in instructional decision making. *Reading today*, 1-8. http://www.reading.org/publications/rty/archives 02decevidente.html. Date of access: 11 December, 2004.

FEUER, M. 2002, February. *The logic and the basic principles of scientific based research.* Presentation given at a seminar on the use of scientifically based research in education, Washington, DC. http://www.ed.gov/print/nclb/methods/whatworks/research/page_pg4.html. Date of access: 16 December, 2003.

FLINDERS, D. J. 2003. Qualitative research in the foreseeable future: No study left behind? *Journal of curriculum and supervision, 18*(4):380-390.

FRICK, R.W. 1996. The appropriate use of null hypothesis testing. *Psychological methods*, 1:379-390.

GLASER, D.N. 1999. The controversy of significance testing: Misconceptions and alternatives. *American journal of critical care*, 8(5):291-296.

GLASS, G.V. 1976. Primary, secondary and meta-analysis of research. *Educational researcher,* 5:3-8.

GLASS, G.V., McGAW, B. and SMITH, M.L. 1981. *Meta-analysis in social research*. London: Sage.

GREENWALD, A.G., GONZALEZ, R., HARRIS, R.J. & GUTHRIE, D. 1996. Effect sizes and p-values: What should be reported and what should be replicated? *Psychophysiology,* 33:175-183.

GROSSEN, B. 1996. Making research serve the profession. *American Educator, 20*(3):7-8, 22-27.

GUPTA, A. 2004. Effects of on-site reading clinical tutoring on children's performance. *The reading matrix*, 4(2):54-62.

HAGAN, R.L. 1997. In praise of the null hypothesis statistical test. *American psychologist*, 52:15-24.

HARRIS, R.J. 1997a. Reforming significance testing via three-valued logic. *In* Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

HARRIS, R.J. 1997b. Significance tests have their place. *Psychological science,* 8:8-11.

HATCH, E. & LAZARATON, A. 1991. The research manual: Design and statistics for applied linguistics. Rowley, MA: Newbury House.

HAYES, A.F. 1998. Reconnecting data analysis and research design: Who needs a confidence interval? *Behavioral and brain sciences, 21:2-3-204.*

HAYS, W.L. 1981. Statistics. 3rd ed. New York : Holt, Rinehart Winston.

HERMAN, F. 2003. Differential effects of reading and memorization of paired associates on vocabulary acquisition in adult learners of English as a second language. Indiana University of Pennsylvania.

HINKLE, D.E. & OLIVIER, J.D. 1983. How large should the sample be? A question with no simple answer? *Educational and psychological measurement,* 43l (4):1051-1060.

HINKLE, D.E., WIERSMA, W. & JURS, S.G. 1994. Applied statistics for the behavioural sciences. 3rd ed. Boston: Houghton Mifflin Company.

HOPKINS, W.G. 1997. New view of statistics. http://www.sports.ci.org/resource/stats/effectmag.html. Date of access: 23 August, 2002.

HUBBARD, R., PARSA, A.R. & LUTHY, M.R. 1997. The spread of statistical significance testing in psychology: The case of the journal of applied psychology. *Theory and psychology*, 7:545-554.

HUBERTY, C. 1993. Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of experimental education*, 61:317-333.

HUBERTY, C.J. 2002. 'A history of effect size indices'. *Educational and psychological measurement*, 62(2):227-240.

HUNTER, J. E. 1997. Needed: A ban on the significance test. *American psychological society*, *8*:3-7.

INTERNATIONAL READING ASSOCIATION. 2002. What is evidence-based reading instruction? Newark: International Reading Association.

JOHNSON, A. & HOWARD, C.A. 2003. The effects of the accelerated reader program on the reading comprehension of pupils in grades three, four, and five. *The reading matrix*, 3(3):87-96.

KAUFMAN, A.S. 1998, ed. Introduction to the special issue on statistical significance testing. *Research in the schools,* 5(2):1.

KERLINGER, F.N. 1979. Behavioral research: A conceptual approach. New York: Holt.

KESELMAN, H.J., HUBERTY, C.J., LIX, L.M., OLEJNIK, S., CRIBBIE, R.A., DONAHUE, B., KOWALCHUK, R.K., LOWMAN, L.L., PETOSKEY, M.D., KESELMAN, J.C. & LEVIN, J.R. 1998. 'Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyse'. *Review of educational research,* 68(3):350-386.

KIRK, R. E. 1996. Practical significance: A concept whose time has come. *Educational and psychological measurement*, 56:746-759.

KIRK, R.E. 2001. Promoting good statistical practices: Some suggestions. *Educational and psychological measurement*, 61(2):213-218.

KLINE, R.B. 2004. Beyond significance testing. Reforming data analysis methods in behavioural research. Washington, DC: American Psychological Association.

KNAPP, T.R. 1998. Comments on the statistical significance testing articles. *Research in the schools*, 5(2):39-41.

KOTRLIK, J.W. & WILLIAMS, H.A. 2003. The incorporation of effect size in information technology, learning and performance research. *Learning and performance journal*, 21(1):1-5.

KRANTZ, D.H. 1999. The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 94:1372-1381.

KRASHEN, S. 2002. Bilingual education. In B. J. Guzzetti (Ed.) *Literacy in America: An encyclopedia of history, theory, and practice.* Vol. 1, 49-52 pp. Santa Barbara, CA: ABC-CLIO.

LAO, C.Y. & KRASHEN, S. 2000. The impact of popular literature study on literacy development in EFL: More evidence for the power of reading. *System*, 28(2):261-270.

LAUER, P.A. 2004. A policymaker's primer on education research. http://www.ecs.org/html/educationissues/REEsearch/primer/foreword.asp Date of access: 6 August, 2005.

LAWRENCE, C. 2004. Confidence Intervals. Paper presented at Millsaps College, Jackson, Mississippi.

LEVIN, J.R. 1998. What if there were no more bickering about statistical significance tests? *Research in the schools,* 5(2):43-53.

LEVINE, A., FERENZ, O. & REVES, T. 2000. EFL Academic reading and modern technology: How can we turn our students into independent critical readers? Bar Ilan University, Israel.

LOFTUS, G. R. 1991. On the tyranny of hypothesis testing in the social sciences. *Contemporary psychology*, 36:102-104.

MARZANO, R. J., PICKERING, D. J., & POLLOCK, J. E. 2001. Classroom instruction that works: Research-based strategies for increasing student achievement. Alexandria, VA: Association for supervision and curriculum development.

MAXWELL, S.E., CAMP,. C.J. & AVERY, R.D. 1981. Measures of strength of association: A comparative examination. *Journal of applied psychology*, 66:525-534.

McEWAN, E.K. & McEWAN, P.J. 2003. Making sense of research: What's good, what's not, and how to tell the difference. Thousand Oaks, C.A.: Sage Publications.

McGRAW, K.O. 1991. Problems with the BESD: a comment on Rosenthal's "How are we doing in soft psychology". *American psychologist,* 46:1084-6.

McGRAW, K.O. and WONG, S.P. 1992. 'A Common Language Effect Size Statistic'. *Psychological bulletin,* 111:361-265.

McLEAN, J.E. & ERNEST, J.M. 1998. The role of statistical significance testing in educational research. *Research in the schools*, 5(2):15-22.

MEEHL, P.E. 1997. The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. *In* L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* 391-423 pp. Hillsdale, NJ: Erlbaum.

MILLER, L.E. 1998. Appropriate analysis. *Journal of agricultural education*, 39(2):1-10.

MOHER, D., SCHULZ, K.F. & ALTMAN, D.G. 2001. The Consort statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*, 357:1191-1194.

MOORE, D.S. 1995. The basic practice of statistics. New York: Freeman.

MURRAY, L.W. & DOSSER, D.A. 1987. How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of counselling psychology*, 34(10):68-72.

NATIONAL CLEARINGHOUSE FOR COMPREHENSIVE SCHOOL REFORM. 2001. Taking stock: Lessons on comprehensive school reform from policy, practice, and research. *Benchmarks*, 2:1-11.

NATIONAL INSTITUTE OF CHILD HEALTH AND HUMAN DEVELOPMENT (NICHD). 2000. Report of the national reading panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Reports of the subgroups (N1H Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.

NATIONAL READING PANEL. 2000. Report of the National Reading Panel: Teaching students to read: An evidence/bases assessment of the scientific research literature on reading and its implications for reading

instruction: Reports of the subgroups. Bethesda, MD: National Institute of child health and human development, National institutes of health.

NATIONAL RIGHT TO READ FOUNDATION. 2003. California state educators examine evidence-based reading. Www.nrrf.org/ca-state-pri-1-15-03.htm. Date of access: 11 November, 2005.

NESTER, M.R. 1996. An applied statistician's creed. *Applied statistics*, 45:401-410.

NICKERSON, R.S. 2000. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological methods, 5:*241-301.

NIKOLOVA, O.R. 2002. Effects of students' participation in authoring of multimedia materials on student acquisition of vocabulary. *Journal of language learning and technology,* 6(1):100-122.

NIX, T.W. & BARNETTE, J.J. 1998. A review of hypothesis testing revisited: Rejoinder to Thompson, Knapp and Levin. *Research in the schools*, 5(2):55-57.

NO CHILD LEFT BEHIND, Act of 2002. Pub.L. No. 107-110, 115 Stat. 1425 2002. http://www.ed.gov/legislation/ESEA02/. Date of access: 17 December, 2003.

NUNAN, D. 1991. Methods in second language classroom-oriented research. *Studies in second language acquisition*, 13:249-274.

NUNES, T., BRYANT, P. & OLSSON, J. 2003. Learning morphological and phonological spelling rules: An intervention study. *Scientific studies of reading*, 7(3):289-307.

OLJENIK, S. & ALGINA, J. 2000. Measures of effect size for comparative studies: Applications, interpretations and limitations. *Contemporary educational psychology, 25:* 241-286.

PATTERN, M.L. 2002. Understanding research methods: An overview of the essentials. 3rd ed. Los Angeles: Pyrczak Publishing.

PATTON, M.W. 1990. Qualitative evaluation and research methods. 2nd ed. Newbury park, CA: Sage.

PEARSON, K. 1901. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London*, 195:1-47.

PEDHAZUR, E.J. & SCHMELKIN. O.P. 1991. Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

PORTILLO, M.T. 2001. Statistical significance tests and effect magnitude measures within quantative research manuscripts published in the Journal of Agricultural Education during 1996-2000. Paper presented at the 28th annual national agricultural education research conference. Oklahoma State University.

PRESSLEY, M. 2003. A few things reading educators should know about instructional experiments. *Reading Teacher*, 57(1). http://www.readingonline.org/articles/art_index.asp?HREF=RT/9-03_column/index.html. Date of access: 11 November, 2005.

PRESSLEY, M. 2006. What the future of reading research could be. Paper presented at the international reading association's reading research 2006, Chicago, Illinois.

PRUZEK, R.M. 1997. An introduction to Bayesian inference and its application. *In* L.L. Harlow, S.A. Mulaik & J.J. Steiger (Eds.). *What if there were no significance tests?* 187-318 pp. Hillsdale, NJ: Erlbaum.

PUBLICATION MANUAL OF THE AMERICAN PSYCHOLOGICAL ASSOCIATION. (1994). 4th ed. Washington, DC: American Psychological Association.

PURCELL-GATES. V. 2000. The role of qualitative and ethnographic research in educational policy. *Reading Online*, 4(1). http://www.readingonline.org/articles/purcell-gates/. Date of access: 9 October, 2005.

RASEKH, Z.E. & RANJBARY, R. 2003. Metacognitive strategy training for vocabulary learning. Iran University of Science and Technology. TESL EJ.

REYNA, V. 2002. 'What is scientifically based evidence? What is its logic? Presentation given at a seminar on the use of scientifically based research in education, Washington, DC. http://www.ed.gov/print/nclb/methods/whatworks/research/page_pg3.html. Date of access: 16 December 2003.

ROSENTHAL, R, & RUBIN, D.B. 1982. A simple, general purpose display of magnitude of experimental effect. *Journal of educational psychology*, 74: 166-169.

ROSNOW, R.L. & ROSENTHAL, R. 1996. Compating contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological methods*, 1:331-340.

ROSNOW, R.L., ROSENTHAL, R. & RUBIN, D.B. 2000. Contrasts and correlations in effect size estimation. *Psychological science*, 11:446-453.

ROSSI, J.S. 1997. A case study in the failure of psychology as a cumulative science; The spontaneous recovery of verbal learning. *In* L.L. Harlow, L.L., Mulaik, S.A. & Steiger, J.H. (Eds.). *What if there were no significance test ?* 175-197 pp. Hillsdale, NJ: Erlbaum.

ROZEBOOM, W.W. 1997. Good science is abductive, not hypothetico-deductive. (In Harlow, L.L., Mulaik, S.A. & Steiger, J.H. (Eds.). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.)

RSA DoE (Department of Education). 2001. National report on systemic evaluation foundation phase mainstream. Pretoria: Chief Directorate: Quality Assurance, Department of Education.

RUSSEK, B.E. & WEINBERG, S.L. 1993. Mixed methods in a study of implementation of technology-based materials in the elementary classroom. *Evaluation an program planning*, 16(2):131-142.

SCHMIDT, F.L. & HUNTER, J.E. 1997. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. (In Harlow, L.L., Mulaik, S.A. & Steiger, J.H. (Eds.). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.)

SCHMIDT, F.L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological methods,* 1:115-129.

SELIGER, H.W. & SHOHAMY, E. 1989. Second language research methods. Oxford: Oxford University Press.

SERLIN, R. C., & LAPSLEY, D. K. 1985. Rationality in psychological research. *American psychologist*, 40:73-83.

SERLIN, R.C. 1993. Confidence intervals and the scientific method: A case for holm on the range. *Journal of experimental education*, 61:350-360.

SHAVELSON, R. J., & TOWNE, L. (Eds.). 2002. *Scientific research in education.* Washington, DC: National Academy Press.

SLAVIN, R. E. 2003. A reader's guide to scientifically based research. *Educational leadership, 60*(5):12-16. http://www.ascd.org/publications/ed_lead/200302/slavin.html. Date of access: 16 December, 2003.

SLAVIN, R.E. 2003. A reader's guide to scientifically based research. *Educational Leadership, 60*(5):12-16. http://www.ascd.org/cms/objectlib/ascdframeset/index.cfm?publication=http://www.ascd.org/publications/ed_lead/200302/toc.html. Date of access: 13 May, 2004.

SNOW, C., BURNS, M.S. & GIFFIN, P. (Eds.). 1998. Preventing reading difficulties in young children. Washington, DC: National Academy Press.

STAHL, S.A., DUFFY-HESTER, A.M. & STAHL, K.A. 1998. Everything you wanted to know about phonics (but were afraid to ask). *Reading research quarterly,* 33:338-355.

STAKHNEVICH, J. 2002. Reading on the Web: Implications for ESL Professionals. *Journal: The reading matrix*, 2(2):1-30.

STANOVICH, P. J., & STANOVICH, K. E. 2003. Using research and reason in education. Jessup, MD: National Institute for Literacy http://www.nifl.gov/partnershipforreading/publications/k-3.html Date of access: 13 May, 2004.

STATSOFT, 2005. An overview of Statistica. :http://www.statsoft,com/uniquefeatures/general.htm. Date of access: 18 February, 2006.

STATSOFT. 2005. Electronic textbook Statsoft www.statsoft.com/textbook/stdisfit.html. 6 January, 2006.

STEIGER, J.H. & FOULADI, R.T. 1997. Noncentrality interval estimation and the evaluation of statistical models. *In* Harlow, L.L., Mulaik, S.A. & Steiger, J.H. (Eds.). *What if there were no significance tests?* Mahwah, N.J.: Lawrence Erlbaum Associates.

STEYN, H.S. 2000. Practical significance of the difference in means. *Journal of industry psychology, 26(3):1-3.*

STEYN, H.S. (jr.). 2005. Handleiding vir die bepaling van effekgrootte-indekse en praktiese betekenisvolheid http://www.puk.ac.fakulteite/natuur/skd/index.html Noordwes-Universiteit (Potchefstroomkampus), Potchefstroom. 17 January, 2006.

STRAHAN, R.F. 1991. Remarks on the binomial effect size display. *American psychologist,* 46:1083-4.

STRAUSS, A. & CORBIN, J. 1990. Basics of qualitative research: Grounded theory procedures and techniques. Newbury Park, CA: Sage.

SVYANTEK, D.J. & EKEBERG, S.E. 1995. The earth is round (so we can probably get there from here). *American psychologist*, 50:1101.

SYMONS, S., MACLATCHY-GAUDET, H., STONE, T.D. & LEE REYNOLDS, P. 2001. Strategy instruction for elementary students searching informational text. *Scientific studies of reading*, 5(1):1-33.

THALHEIMER, W. & COOK, S. 2002. How to calculate effect sizes from published research articles: A simplified methodology http://work-learning.com/effect-sizes.htm. Date of access: 7 February, 2006.

THE NATIONAL RIGHT TO READ FOUNDATION. 2003. California State educators examine evidence-based reading. Instruction and teacher quality at new reading faculty forum: evidence-based research in reading: Implications and change state pri-1-15-03.htm. 12 August, 2004.

THOMAS, J.R., SALAZAR, W. & LANDERS, D.M. 1991. What is missing in p<.05? Effect size. *Research quarterly for exercise and sport,* 62(3): 344-348.

THOMPSON, B. & SNYDER, P.A. 1997. Statistical significance testing practices in The journal of experimental education. *The journal of experimental education*, 66(1):75-79.

THOMPSON, B. & SNYDER, P.A. 1998. Statistical significance and reliability analysis in recent *JCD* research articles. *Journal of counselling and development*, 76:436-441.

THOMPSON, B. & SNYDER, P.A. 1998. Statistical significance testing practices in the Journal of Experimental Education. *Journal of experimental education*, 66:75-83.

THOMPSON, B. 1987. The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. *Paper presented at the Annual Meeting of the American educational Research Association.* Washington, DC.

THOMPSON, B. 1994. The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of personality*, 62:157-176.

THOMPSON, B. 1995. Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. *Educational and psychological measurement,* 55:84-94.

THOMPSON, B. 1996. AERA editorial polices regarding statistical significance testing: Three suggested reforms. *Educational researcher,* 25(2):26-30.

THOMPSON, B. 1998. In praise of brilliance: Where that praise really belongs. *American psychologist,* 53:799-800.

THOMPSON, B. 1999a. If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & psychology*, 9:167-183.

THOMPSON, B.. 1999b. Improving research clarity and usefulness with effect size indices as supplements to statistical significance tests. *Exceptional children*, 65:329-337.

THOMPSON, B. 1999c. Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. *Educational psychology review*, 11:157-169.

TRYON, W.W. 1998. The inscrutable null hypothesis. *American psychologist*, 53:796.

TUKEY, J.W. 1991. The philosophy of multiple comparisons. *Statistical science*, 6:1 00-116.

U.S. Department of Education. 2002, June 25. Statement of Governor J. Whitehurst before the senate committee on health, education, labor and pensions [Press release]. http://www.ed.gov/news/speeches/2002/06/06252002.html?exp=0. Date pf access: 16 December, 2003.

UYS, C. & DREYER, C. 2005. The effect of knowledge of high frequency words on the reading ability of learner in the foundation phase. Paper

represented at the EASA Conference, North-West University, Potchefstroom. (Potchefstroom Campus.)

VACHA-HAASE, T. & NILSSON, J.E. 1998, April. Statistical significance reporting: Current trends and uses in MECD. *Measurement and evaluation in counselling and development,* 31(10):46-57.

VACHA-HAASE, T. 2001. Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and psychological measurement,* 61(2):291-224.

VAN LJZENDOOM, M.H. 1997. Meta-analysis in early childhood education: Progress and problems. *In* Spodek, B., Pellegrini, A.D. & Saracho, O.N. (Eds.), *Issues in early childhood education yearbook in early childhood education.* New York: Teachers College Press.

VAN WYK, A.L. 2001. The development and implementation of an English language and literature programme for low-proficiency tertiary learners. Unpublished thesis. Bloemfontein: University of the Free state.

VASQUES, L.M., GANGSTEAD, S.K. & HENSON, R.K. 2000. Understanding and interpreting effect size measures in general linear model analysis. *Paper presented at the Annual Meeting of the Southwest Education Research Association.*

VERNON, D.T.A. & BLAKE, R.L. 1993. Does problem-based learning work? A meta-analysis of evaluative research. *Academic medicine,* 68:550-563.

WAGNER, D.A. 2000. EFA 2000 thematic study on literacy and adult education: Paper presented at the World Education Forum, Dakar. Philadelphia; International Literacy Institute.

WHAT WORKS CLEARINGHOUSE. 2006a. Study design classification. :http://whatworks.ed.gov/reviewprocess/studydesignclass.pdf. Date of access: 5 March, 2006.

WHAT WORKS CLEARINGHOUSE. 2006b. Study review standards. http://www.whatworks.ed.gov/reviewprocess/study_standards_final.pdf. Date of access: 5 March, 2006.

WIKIPEDIA. 2006. Statistics. http://en.wikipedia.org/wiki/Statistical#Statistical techniques.

WILKINSON, L. & The American Psychological Association. The Task Force on statistical inference. 1999. Statistical methods in psychology journals: Guidelines and explanations. *American psychologist,* 54:594-604.

WILLIAMS, H.A. 2003. A mediated hierarchical regression analysis of factors related to research productivity of human resource development postsecondary faculty Doctoral Dissertation. Louisiana State University, 2003. http://etd01.lnx390.lsu.edu:80585/docs/available/etd - 0326103-212409/. Date of access: 9 April, 2003.

YATES, F. 1951. The influences of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, 46:19-34.

ZAKZANIS, K.K. 1998. Brain is related to behavior ($p<.05$). *Journal of clinical and experimental neuropsychology*, 20:419-427.

APPENDIX A: SUBMISSION GUIDELINES (THE READING MATRIX)

Information for Authors

The Reading Matrix is seeking submissions of previously unpublished manuscripts on any topic related to the teaching and learning of reading, the application of technology to literacy instruction, and issues in second language learning and teaching. Some areas of interest are personal characteristics and attitudes of readers, L2 learners' reading processes, readers' responses to texts, assessment/evaluation of reading comprehension and proficiency, vocabulary acquisition, reading and computer-assisted instruction, and any other topic clearly relevant to L2 pedagogy and research. We welcome both practical and research focused articles. Articles should have a clear focus and be written so that they are accessible to a broad audience of reading and language educators, including those individuals who may not be familiar with the particular subject matter addressed in the article. Manuscripts are being solicited in the following categories:

Articles
Articles should report on original research or present an original framework that links previous research, educational theory, and teaching practices. Full-length articles should be no more than 7500 words in length and should include an abstract of no more than 200 words. We encourage articles that take advantage of the electronic format by including hypermedia links to multimedia material both within and outside the article. All article manuscripts submitted to *The Reading Matrix* go through an internal review first which takes 1-2 weeks. If the article meets the basic requirements, they are then sent out for external double-blind review from experts in the field, either from either from the journal's editorial board or from our larger list of reviewers. This second review process takes 1-2 months. Following the external review, the authors are sent copies of the external reviewers' comments and are notified as to the decision (accept as is, accept pending changes, revise and resubmit, or reject).

Submission Guidelines for Articles

Please list the names, institutions, e-mail addresses, and if applicable, World Wide Web addresses (URL(s)), of all authors. Also include a brief biographical statement (maximum 50 words, in sentence format) for each author. (This information will be temporarily removed when the articles are distributed for blind review).

APPENDIX B: SUBMISSION GUIDELINES (LANGUAGE LEARNING AND TECHNOLOGY)

LANGUAGE LEARNING & TECHNOLOGY
a refereed journal for second and foreign language educators

LLT RESEARCH GUIDELINES for QUANTITATIVE and QUALITATIVE RESEARCH

Guidelines for Articles Reporting on Quantitative Research

The *LLT* editors recommend that authors of manuscripts based on quantitative research consider the general guidelines outlined in Chapter 1 of the *Publication Manual of the American Psychological Association* (4th edition. 1995. Washington, D.C.: American Psychological Association, pp. 7-22).

In particular, a quantitative research report should generally include the following sections:

An **Introduction** that

- states the problem to be investigated
- contextualizes the research by describing the underlying theoretical framework and reviewing previous studies
- defines the variables and research hypotheses

A **Method Section** that describes

- the participants (e.g., demographics, selection criteria, and group assignment)
- the materials (e.g., task[s], equipment, instruments, including a discussion of their validity and reliability, if appropriate)
- the procedures employed in the study such as treatment(s)

A **Results Section** that includes

- graphs and tables that help to present and explain the results
- descriptive and inferential statistics used to analyze the data, including the following:
    - name of the statistic used and in the case of an uncommon statistical procedure, a reference to a discussion of the procedure
    - statistical significance of the results obtained
    - measures of effect sizes
    - how all necessary assumptions were met

A **Discussion Section** that includes

- an interpretation of the results
- an explanation of the results, including alternative explanations when appropriate
- a statement relating the results obtained in the study to original hypotheses
- theoretical implications

- limitations of the study

A **Conclusion** that includes

- general implications of the study
- suggestions for further research

References

Appendices of instrument(s) used

Guidelines for Articles Reporting on Qualitative Research

The *LLT* editors recommend that authors of manuscripts based on qualitative research generally include the following sections in their articles:

- Statement of the **research question** examined in the study.
- Description of the **theoretical framework(s)** underlying the research question.
- Description of the **methodological tradition** in which the study was conducted.
- Relationship between the study and **previous work** in the area under investigation.
- Detailed description of the **participants** and **research site**.
- Detailed description of **data collection and analysis** procedures.
- Report of **findings**.
- **Limitations** of the study.
- **Implication(s)** of the study.

APPENDIX C: SUBMISSION GUIDELINES (SCIENTIFIC STUDIES OF READING)

Society for the Scientific Study of Reading

SSSR

Home

Join/Renew Now

The Journal

Members Section

Conferences

About SSSR

SSSR Links

Contact Us

CONTRIBUTOR INFORMATION

Content: *Scientific Studies of Reading (SSR)* is the official journal of the Society for the Scientific Study of Reading and publishes original empirical investigations of all aspects of reading and its related areas, and occasionally, scholarly reviews of the literature, papers focused on theory development, and discussions of social policy issues. Commentary or criticism on topics pertinent to the journal's concerns will also be considered for publication.

Manuscript Submission: Prepare your manuscript in Word or another popular word-processing program and submit one electronic copy by e-mail to Editor Elect Charles Hulme at ch1@york.ac.uk. Follow the instructions for removing identifying information provided on this page under the Blind Review section.

Only articles published in English will be considered. Prepare manuscripts according to the *Publication Manual of the American Psychological Association* (APA; 5th ed., 2001). Type all components of the manuscript double-spaced, including title page, abstract, text, quotes, acknowledgments, references, appendixes, tables, figure captions, and footnotes. An abstract of 100-150 words should be typed on a separate page. Authors must use nonsexist language in their articles as specified in the "Guidelines to Reduce Bias in Language" on pp. 61-76 of the *APA Manual* (5th ed.). One photocopy of illustrations and the original illustrations should accompany the manuscript. Illustrations should also be sent as an electronic file to the editor, along with the manuscript. All manuscripts submitted will be acknowledged promptly. Authors should keep a copy of their manuscripts to guard against loss.

Blind Review: To facilitate anonymous review, only the article title should appear on the first page of the manuscript. An attached cover page must contain the title; authorship; authors' affiliations; any statements of credit or research support; and mailing addresses, phone numbers, fax numbers, and e-mail addresses (if available) of the authors. Every effort should be made by the authors to see that

the manuscript itself contains no clues to their identities.

Permissions: Authors are responsible for all statements made in their work and for obtaining permissions from copyright owners to reprint or adapt a table or figure or to reprint a quotation of 500 words or more. Authors should write to original author(s) and publisher to request nonexclusive world rights in all languages to use the material in future editions. Provide copies of all permissions and credit lines obtained.

Regulations: In a cover letter, authors should state that the findings reported in the manuscript are original and have not been published previously and that the manuscript is not being simultaneously submitted elsewhere. Authors should also state that they have complied with American Psychological Association ethical standards in the treatment of their samples.

Production Notes: After a manuscript is accepted for publication, its author is asked to provide a computer disk containing the manuscript file. Files are copyedited and typeset into page proofs. Authors read proofs to correct errors and answer editors' queries. Correction of typographical errors is made without charge; other alterations are to be prepaid by authors. Authors may order reprints of their article only when they return page proofs.

APPENDIX D: SUBMISSION GUIDELINES (SYSTEM)

Submission of Papers

Submission of a paper implies that it has not been published previously, that it is not under consideration for publication elsewhere, and that if accepted it will not be published elsewhere in the same form, in English or in any other language, without the written consent of the publisher. If a related article is being published by the author elsewhere, that fact should be indicated by the contributor.

Manuscript Preparation
Format: Manuscripts must be double-spaced, 12 or 10 pt,with wide margins on one side . Authors should consult a recent issue of the journal for style at http://www.sciencedirect.com/science/journal/0346251X. Follow this order when typing manuscripts: Abstracts, Keywords, Main Text, Acknowledgements, Appendix, References, Figure Captions and then Tables. Do not import the Figures or Tables in to your text. The corresponding author should be identified with an asterisk and footnote. All other footnotes (except for table footnotes) should be identified with superscript Arabic numbers. Material cited at length in the text should be indented.

The Editor reserves the right to adjust style to certain standards of uniformity. The corresponding author should be identified (include a Fax number and E-mail address). Full postal addresses must be given for all authors. Authors should retain a copy of their manuscript since we cannot accept responsibility for damage of loss of papers.

Electronic submission is preferred. Please submit an electronic copy of the manuscript in both PDF and a format compatible with Word or Wordperfect (e.g. RTF (rich text format)) to the Editor, Norman F. Davies, at norda@isk.liu.se Important notice: the electronic manuscript should be **anonymous**: authors should not identify themselves in the electronic manuscript, and are requested to include their name, the title of the paper, vitae, and their address/affiliation in the e-mail message.

Full details of electronic submission and formats can be obtained from http://authors.elsevier.com

Paper length: Manuscripts should not normally exceed 5000 words.

Abstracts: Each article should include an abstract of between 150 and 200 words which succinctly presents the content of the article.

Keywords: Each article should be supplied with a maximum of 10 keywords or phrases, which are for indexing purposes, and should conform to those used in *Linguistics and Language Behaviour Abstracts*, as far as possible.

References: All publications cited in the text should be presented in a list of references following the text of the manuscript. In the text refer to the author's name (without initials) a year of publication (e.g. "Since Peterson (1993) has shown that..." or "This is in the agreement with results obtained later (Kramer, 1994)"). For three or more authors use the first author followed by "et al"., in the text. The list of references should be arranged alphabetically by authors' names. The manuscript should be carefully checked to ensure that the spelling of authors' names and dates are exactly the same in the text as in the reference list. References should be given in the following form: Allen, S.D., 1991. Ability grouping research review: what do they say about grouping and the gifted? Educational Leadership 48 (6), 60-65. Barnes, D., 1976. From Communication to Curriculum. Penguin, New York. Gumperz, J.J., 1975. On the ethnology of linguistic change. In: Bright, W. (Ed.). Sociolinguistics: Proceedings of the UCLA Sociolinguistics Conference 1964. Mouton, The Hague, pp. 27-38. Porter, P., 1986. How learners talk to each other: input and interaction in task-centered discussions. In: Day, R. (Ed.), Talking to Learn: Conversation in Second Language Acquisition. Newbury House, Rowley, MA, pp. 200-222. The titles of journals should be abbreviated, and should follow the convention of *Bibliographie Linguistique*.

Illustrations: All illustrations should be provided in camera-ready form, suitable for

reproduction (which may include reduction) without retouching. Photographs, charts and diagrams are all to be referred to as "Figures(s)" and should be numbered consecutively in the order to which they are referred. They should accompany the manuscript, but should not be included within the text. All illustrations should be clearly named with the figure number and the author's name. All figures are to have a caption.

Line drawings: Good quality printouts on white paper produced in black ink are required. All lettering, graph lines and points on graphs should be sufficiently large and bold to permit reproduction when the diagram has been reduced to a size suitable for inclusion in the journal. Dye-line prints of photocopies are not suitable for reproduction. Do not use any type of shading on computer-generated illustrations.

Photographs: Original photographs must be supplied as they are to be reproduced (e.g. black and white or colour). If necessary, a scale should be marked on the photograph. Please note that photocopies of photographs are not acceptable.

Colour: Authors will be charged for colour at current printing costs.

Tables: Tables should be numbered consecutively and given a suitable caption and each table typed on a separate sheet. Footnotes to tables should be typed below the table and should be referred to by superscript lowercase letters. No vertical rules should be used. Tables should not duplicate results presented elsewhere in the manuscript, (e.g. in graphs).

APPENDIX E: SUBMISSION GUIDELINES (TEACHING ENGLISH AS A SECOND OR FOREIGN LANGUAGE)

TESL-EJ Teaching English as a Second or Foreign Language

Submission Procedures

TESL-EJ publishes original articles in the research and practice of English as a second or foreign language. TESL-EJ welcomes studies in ESL/EFL pedagogy, second language acquisition, language assessment, applied socio- and psycholinguistics, and other related areas, for quarterly publication.

Because of the large number of submissions, all article proposals must be pre-screened. Please use our screening form if you wish to submit an article.

Once you have received a confirmation number allowing you to submit an article, please follow these directions carefully:

Articles, reviews, or forum discussion should conform generally to the American Psychological Association (5th Edition) format, with the following adjustments made for computer-based dissemination:

Text should be prepared in Microsoft Word or in RTF. Please put all figures and tables in place in the text (rather than at the end). Number and title all figures and tables.

Manuscripts may be sent by e-mail attachment. Please notify the editor in advance for transfer of large sound, graphics, or video files.

Notes should be endnotes, and kept to a minimum. Endnotes should be numbered, using square brackets, consecutively

within the text. For example:

...this unusual methodology [1]...
These notes will link to the actual endnotes.

All initial submissions should be sent by e-mail to the appropriate editor (see below). The editors reserve the right to reject poorly edited or improperly formatted manuscripts.

Articles: Full-length articles should include an abstract no more than 150 words. Before accepting any article for publication, the editor shall solicit recommendations by two qualified reviewers. These reviewers will be members of the TESL-EJ Editorial Board, or other qualified experts. The names of all readers will be disseminated annually, in the last quarter of the publication year.

Classroom Focus: Classroom Focus articles are reviewed in the same manner as main articles.

You will receive an e-mail address to use for your submission once you have completed the screening process.

Other TESL-EJ sections:

Book Reviews: Please read the Book Review Policies.

See Books Received for titles available for review.

Media Reviews: Please read the Media Review Policies.

Forum and Discussion: The Forum, a vehicle for discussion of topics of interest, also invites contributions. Query:

(position vacant), Forum Editor
<forum@tesl-ej.org>

On the Internet: Query Editor Vance Stevens. You can find contact information at his website:
http://www.vancestevens.com/vance.htm

APPENDIX F: SUBMISSION GUIDELINES (THE SOUTH AFRICAN JOURNAL FOR LANGUAGE TEACHING)

Journal for Language Teaching

Author Guidelines

Any researcher in the field of language teaching is free to submit papers to the Journal for Language Teaching. Papers of high standard from outside South Africa which prove to be of general importance to the teaching of languages, may be submitted to the Journal.

Selection of papers

A strict policy of selection is adhered to. All papers submitted for publication in the Journal for Language Teaching are sent to at least two peer reviewers for blind peer reviewing. Reports on these reviews are compiled by the editor and then sent to the author(s). If one or both of the reviews are negative, the author(s) may resubmit a revised copy. The revised copy is then again sent to one or two reviewers, depending on the first reports.

A list of relevant questions (attached) is always sent to the reviewers. Reviewers may also make additional comments and may comment on the paper itself.

Members of the Editorial Board and peer reviewers

All the members of the Editorial Board are chosen because of their research work and the quality and quantity of their publications, locally and abroad.

The Journal also has a list of additional peer reviewers that is constantly extended – all of them specialists in their respective fields of language teaching.

Language policy

The Editorial Board has no language restriction or fixed quota for papers in the Journal. Nearly all papers are in English and some in Afrikaans. If a paper is written in Afrikaans (or in any other South African language except for English), it must be supported by a substantial abstract in English, with an English heading and with English keywords.

Full address particulars of all contributors to the Journal are published with their

papers.

Submission of manuscripts

The following must be submitted to the Editor:

1.• A letter of submission with

(i) the full address and e-mail particulars of the author(s)and the title of the paper. In the case of more than one author, the name of the author designated to receive the correspondence and to be responsible for the payment of the page fees, must be marked with an *

(ii) titles, names, full addresses and e-mails of three scholars in the same field.

2. A copy of the paper per e-mail, or on a markeddiskette accompanied by a hard copy per snail mail. Full addressand e-mail particulars must also be on the paper. The Abstractmust be on the same document. Keywords must follow theAbstract.

• Electronic documents must be named with the author'ssurname.

• Do not register and do not send per courier.

Submission of a paper is taken to imply that it has not beenpublished elsewhere and that it is not under considerationelsewhere, and also that it is language edited. All manuscripts are sent to two scholars for blind peer reviewby the Editor. Further information will be sent to the author(s) after thereviewing.

Page fees

SAALT Members: R80 per printed page

Non-members: R110 per printed page

Word processing

Use `Normal'. No pre-programmed styles.

No capital letters for headings. Use bold and italics.

No automatic bullets and numbering.

Length: 5000–6000 words.

Abstract:

All papers must have an English Abstract of approximately 150words.

An additional Abstract in another language may be included.

References

References in the text

When a page is relevant: Smith (1997: 14); Smith (1997:18–20); (Smith, 1997: 50) or (Smith 1997: 50).

When a whole work is relevant: Smith (1997); (Smith, 1997) or(Smith 1997).

Reference List

All works referred to in the text must be listed in the Reference List and vice versa.

Single-author books / Net een outeur

Turke, S. 1996. *Life on the Screen*. London: Weidenfeld & Nicholson.

Dual-author books / Twee outeurs

Weitsman, E. & Miles, M.M. 1995. *Computer programmes for Qualitative Data Analysis*. Calif: Sage.

Edited books

Kemmis, S. & McTaggart, R. (eds.) 1994. *Instructional Effectiveness of Video Media* (3rd ed.). Hillsdale: Lawrence Erlbaum.

Articles in books

Thompson, I. 1995. Assessment of foreign language comprehension. Pp. 14–31 in D.J. Mendelsohn & J. Rubin(eds.), *A guide for the Teaching of Second Language Listening*. San Diego: Dominie Press.

Articles in journals / Artikels in tydskrifte

Savery, J.R. & Duffy, T.M. 1995. Problem based learning. *Educational Technology* 35: 31–38.

Unpublished

Park-Oh, Y.Y. 1994. Self-regulated strategy training in secondlanguage reading. Unpublished doctoral thesis, University ofAlabama.
Newspaper articles
Sunday Express 1998. 25 November: 12.
Internet & CD-Rom:
Smit, R. 1999. Rekenaarmatige verwysings. Journalfor Language Teaching / Tydskrif vir Taalonderrig 33/2:165–173.

Submission Preparation Checklist (All items required)

- The submission has not been previously published nor is it before another journal for consideration; or an explanation has been provided in Comments to the Editor.

Copyright Notice

Privacy Statement

The names and email addresses entered in this journal site will be used exclusively for the stated purposes of this journal and will not be made available for any other purpose or to any other party.