Reporting effect sizes as a supplement to statistical significance testing in reading instruction studies

C. Nel
North-West University
School of Education
South Africa
e-mail: carisma.nel@nwu.ac.za

H. S. Steyn
North-West University
Statistical Consultation Services
South Africa
e-mail: Faans.Steyn@nwu.ac.za

J. Brits
Orbit FET College
South Africa

Abstract
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. Behind the concept of evidence-based instruction lies the notion of scientifically valid and replicable research that can help educators make effective choices. The purpose of the article is to review the statistical significance reporting practices in reading instruction studies and to provide guidelines for when to calculate and report effect sizes in educational research. A review of six readily accessible (online) and accredited 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 measures when documenting their results.

INTRODUCTION
Nearly 40 per cent of fourth grade students in the USA are unable to read at grade level (International Reading Association 2002) while the results of a national evaluation of Grade three learners in South Africa indicate that the average scores obtained by learners across all the provinces, for reading comprehension, was 39 per cent (RSA DoE 2001, 51). According to the results of the 2006 PIRLS study, the majority of learners, more than half of the English and Afrikaans speaking learners and over
80 per cent of African language speakers in South Africa do not even reach the lowest international benchmark, leaving these learners without basic reading skills and strategies to cope with academic tasks (Howie et al. 2008). In seven southern African countries participating in a regional study of achievement from 1995 to 1998, between 1 per cent and 37 per cent of grade 6 students reached the ‘desirable’ level in reading, while 22–65 per cent were at ‘minimum’ level (Zhang 2006, 581). 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 learners 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, departments of education may find themselves implementing fads or incomplete findings. Using a curriculum based on partial findings will not result in better student reading achievement.

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, and Section 9101 [37]) legislation and the International Reading Association (2002, 1) list the following indicators of effectiveness:

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

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

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.

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.
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.

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.

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

The focus of this article is on one of these indicators, namely the rigorous analysis of data. 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.

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. The criticism of statistical testing is, however, growing fierce.

According to Nunan (1991, 259), many applied linguistic studies 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
Reporting effect sizes as a supplement to statistical significance testing in reading instruction studies

intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless substantial . . .

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; Zakzanis 1998).

These criticisms 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. 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 and 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 purpose of this article is to:

• determine what the state of affairs is with regard to statistical significance reporting practices 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);

• provide guidelines in terms of when the calculation and reporting of effect sizes is necessary and/or advisable in educational research.

METHOD OF RESEARCH

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 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.

Inclusion and exclusion criteria

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.

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.

Identification of studies
Between 16 and 22 March 2002 and then again between 1 and 30 November 2005, six journals that regularly publish articles on reading instruction and that were readily accessible online to practicing English educators in South Africa 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).

The titles and abstracts of a total of 583 studies were reviewed for the purpose of inclusion and analysis in this study. A cut-off date of 5 December 2005 for receipt of the papers was selected on pragmatic grounds. A total number of 16 studies met the inclusion criteria and were selected for final analysis purposes (cf. Results).

Data extraction
Using other literature reviews as a guide (Albanese and Mitchell 1993; Dochy, Segers and Buehl 1999), the researchers 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:

- Author(s) and year of publication;
- Journal in which article was published;
- Data analysis techniques and statistics reported – how was the data analysed?
  What techniques were used and what statistics were reported?

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 mathematical 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 and Buehl 1999).

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; 2001) 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).

**RESULTS AND DISCUSSION**

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 the method of research.

**Studies in The Reading Matrix**

In all of the reviewed studies in *The Reading Matrix* (Bell 2001; Stakhnevich 2002; Johnson and Howard 2003; Gupta 2004) the researchers rely very heavily on the statistical significance of their results and based on this they reject the formulated null hypotheses. The main statistical techniques used by these authors included: t-tests, ANCOVA, factorial ANOVA, and MANOVA. In none of the studies reviewed in this journal did the authors make use of effect size measures to indicate the practical significance of their results.

**Studies in Language Learning and Technology**

Although it is the submission policy of *Language Learning and Technology* for authors to follow the APA guidelines when submitting articles for publication, both of these studies (Al-Seghayer 2001; Nikolova 2002) ignored the requirement to conduct effect size analyses. The statistical techniques used by these authors included the Friedman test, the Wilcoxon matched-pair signed rank test, and MANOVA. 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.
Studies in Scientific Studies of Reading

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 (Symons et al. 2001; Castiglioni-Spalten and Ehri 2003; Nunes et al. 2003; Christensen and Bowey 2005). 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 (e.g., eta-squared or f-squared). It seems as if the calculations are there merely for appearance sake (e.g., Christensen and Bowey 2005). In the study conducted by Nunes et al. (2003, 303) they 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’ (our emphasis). The authors make these conclusions based on only the statistical significance of their results. No effect sizes were calculated in this study. The authors also used LSD tests that do not adjust for familywise error rate as opposed to Tukey tests. The most common statistical techniques used by the researchers included MANOVA, ANOVA and post-hoc Tukey tests.

Studies in System

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. Lao and Krashen (2000) rely exclusively on the statistical significance of the results and do not calculate any effect size measures. It 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. 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. The most common statistical techniques used by these researchers were ANCOVA and t-tests.

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

Although it is the submission policy of Teaching English as a Second or Foreign Language for authors to follow the APA guidelines when submitting articles for publication, all of studied reviewed ignored the requirement to conduct effect size analyses (cf. Levine et al. 2000; Hermann 2003; Rasekh and Ranjbary 2003). The authors primarily used t-tests and a Spearman-rho analysis to analyse their results. The authors relied heavily only on statistical significance testing to either confirm or reject the null hypotheses formulated for the studies. In the study conducted by
Levine et al. (2000), the conclusion reached by the authors 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. The description of the methodology employed in Hermann’s (2003) study seems to be very limited. A quasi-experimental non-randomised design with intact classes constituting a very small sample size (n=17) was used in the study. 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’ (Hermann 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 and 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 or randomly selected from different populations (cf. Hatch and 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 and Ranjbary 2003, 11). This conclusion is based on the statistical significance of the results, which is not relevant due to the lack of randomisation. No effect sizes were calculated in this study.

GUIDELINES: WHY AND WHEN TO CALCULATE AND REPORT EFFECT SIZES IN EDUCATIONAL RESEARCH

When authors communicate the findings of their studies there is often a focus on whether or not some intervention had the intended effect, and less attention to how much of an effect the intervention had. That is, it may be possible to state that some reading intervention increased reading scores more than the usual reading instructional techniques, but it is often more difficult for readers/educators to determine how much of a difference the intervention made. Given that the demand for evidence-based research results using alternatives to statistical significance testing is growing (Stanovich and Stanovich, 2003), perhaps the most commonly recommended supplement to solely interpreting P values is to determine the magnitude of effect, known as ‘effect size’ (Kotrlik and Williams 2003). 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. 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
effect size (Kotrlik and Williams 2003). Effect sizes can, therefore, be applied to any measured outcome in Higher Education Research. It allows us to move beyond the simplistic, ‘Does it work or not?’ to the far more sophisticated, ‘How well does it work in a range of contexts?’

Thompson (2004, 612) states that, ‘the focus of inquiry in psychology and education has now shifted to addressing two questions: (1) how big are detected effects? and (2) are the effects replicable?’ To address the first question, researchers are evaluating the magnitude of effects by considering the context of the research (e.g., higher education instructional effectiveness), and by explicit direct comparison of study results with those in prior related studies (Thompson 2002). As regards the second question, replicability of effects is increasingly being judged by evaluating how stable effects are across a related literature, using meta-analysis (Cumming and Finch 2002). One of the main advantages of using effect size is that when a particular experiment has been replicated, the different effect size estimates from each study can easily be combined to give an overall best estimate of the size of the effect. An important consequence of the capacity of meta-analysis to combine results is that even small studies can make a significant contribution to knowledge. The kind of experiment that can be done by a single lecturer in a lecture room that might involve a total of fewer than 30 students. Unless the effect is huge, a study of this size is most unlikely to get a statistically significant result. According to conventional statistical principles, therefore, the experiment is not worth doing. However, if the results of several such experiments are combined using meta-analysis, the overall result is likely to be highly statistically significant (cf. Thompson 2002). Moreover, it will have the important strengths of being derived from a range of contexts (thus increasing confidence in its generality) and from real-life working practice (thus making it more likely that the policy of a Faculty or University is feasible and can be implemented authentically).

The confusion between effect size and statistical significance seems to stem from misconceptions of what statistical significance testing tells the researcher. 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 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 significance. Statistical significance relies heavily on sample size, while effect size assists in the interpretation of results and makes trivial effects easier to ignore, further assisting researchers/educators to decide whether results are practically significant (Kirk 2001).

The following are cases when the calculation and reporting of effect sizes are highly recommended (cf. Steyn 2005). In addition, reference is made to relevant categories of effect size indexes namely standardised differences to compare two means, relationship between variables, comparison of more than two independent groups of measurements and group overlap:
Case 1
When the scale used to measure differences in means is unknown. For example, in a non-standardised questionnaire where questions are asked using a 4-point Likert scale, and it is not clear what a mean difference of 0,3 means. With standardised scales such as a stanine, the researcher knows what the standard deviation of the scale is and can determine what the differences mean within that context; effect sizes are therefore not necessary. However, if the researcher works with the raw scores before standardisation, the scale is usually not known and the calculation of effect sizes (e.g., a standardised difference in means where the difference is divided by the standard deviation) is, therefore, recommended.

Case 2
Variation in measurements obtained on a specific scale may differ depending on the situation or subjects. For example, the standard deviation on a reading comprehension test of a specifically selected group of students may differ from that of the community at large. Because effect size indexes, like the standardised difference in means, compensate for variation in measurement, the use of effect sizes in this case would be meaningful.

Case 3
Probability samples (e.g., random or stratified) are drawn from the target population and a statistically significant result is obtained, but uncertainty exists with regard to the importance of the differences in means or the relationships that exist. In this case effect size indexes are calculated as a second step once statistical significance has been calculated. If researchers work with big samples, the result is very often statistically significant and effect sizes (e.g., the standardised difference between means when means are compared, or the Pearson correlation coefficient when dealing with linear relationships) are used to indicate whether it is also practically significant.

With the use of small samples, a statistically significant difference (or relationship) is usually an important difference (or relationship) and the calculation of effect size indexes is usually not necessary. However, if a statistically non-significant result is found, the effect size may indicate an important difference or relationship. This can mean one of two things, namely that it is coincidence that the obtained result is so strong, or that a big difference or relationship does exist, but that the sample is too small to prove this. The latter would typically lead to a redesign of the research – bigger samples and more accurate measurements. Small samples are usually used in pilot studies and the use of effect sizes can give an indication of whether the research should be continued or not.
Case 4
When conducting meta-analyses effect sizes are needed to collate the results of the different studies. This is why it is important to report effect sizes even when conducting a small study – it can be used by other authors in future research.

Case 5
When the purpose is to determine the power of a statistical test upon completion of the experiment (i.e., post-hoc), the calculation of effect sizes is necessary. Various kinds of effect sizes are relevant depending on the statistical test. For example, for the power of the t-test, the standardised difference is used.

Case 6
In the planning of a study the sample size that is needed to obtain good power (usually 80%) at a certain significance level can be determined. In order to do this effect sizes are necessary and if this is not known it can be estimated by means of a pilot study. As in Case 5, it depends on the statistical test.

Case 7
When complete surveys (census) are conducted where complete populations are involved, the calculation of effect sizes are usually the only means to determine practical significance. Various effect sizes can be calculated, depending on which kind of significance is required.

CONCLUSION
More than ever, educators are expected to make decisions that guarantee quality instruction. The challenge that confronts educators is the need to view evidence that they read through the lens of their particular educational setting (i.e., school classroom or university lecture hall). What conclusions can be drawn when a new type of instruction produces positive achievement relative to existing instruction? At best, one 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 per cent chance that a difference is due to chance, the finding is often referred to as highly significant even though such a 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
Reporting effect sizes as a supplement to statistical significance testing in reading instruction studies

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.

REFERENCES

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


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...*


