JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 35, NO. 3, PP. 237–248 (1998)

GUEST EDITORIAL

Improving the Interpretation and Reporting of Quantitative Research

Léonie J. Rennie

Science and Mathematics Education Center, Curtin University of Technology, GPO Box U1987, Perth, Western Australia 6845

For many years, researchers have bemoaned the fact that practitioners and policy makers take little notice of their findings. As Shymansky and Kyle (1992 p. 756) put it, "Why does so much effort result in such little apparent benefit?"

Some time ago, Amabile and Stubbs (1982) addressed the research–practice gap. They pointed out that many objective, quantitative findings were not valued by teachers, who considered that most researchers did not understand life in real classrooms, and, closeted in their ivory tower, wrote research reports and journal articles which "consisted of convoluted prose and specialized terminology which takes far too much time to decode into meaningful material" (Stubbs, 1982, p. 25). Teachers had more interest in anecdotal reports, which they found more meaningful and to which they could relate more readily than traditional research methods (Amabile, 1982). At about the same time, White (1984) responded to what he considered unduly pessimistic criticism of the state of educational research with specific examples of how research has changed to focus more on the role of participants in teaching and learning. More recent comment about bringing the research–practice gap in science education also has focused on changing research methodologies to include teachers as collaborators (Krockover & Shepardson, 1995) or teachers as researchers (Pekarek, Krockover, & Shepardson, 1996).

Today, much more of the research in science education, at least, has resemblance to the "anecdotal reports" referred to by Amabile (1982), because, as we shall see, most of the research published in mainstream science education journals in 1996 used qualitative rather than quantitative methodologies (to use a gross oversimplification of research methods). However, the gap between research and practice remains. In his 1997 address, retiring NARST President Tom Koballa quoted teachers' views that science education research was not useful to them. He, too, recommended collaboration with practitioners, but also that "we rethink how science education research is communicated to science teachers" (Koballa, 1997, p. 4) as one way to narrow the research–practice gap.

It is the communication of research results that I wish to address herein, and specifically the communication of quantitative results. We cannot simply dispense with approaches to research that provide quantitative results, because there remain questions to be asked that require

 CCC 0022-4308/98/030237-12

a quantitative answer (Caliendo & Kyle, 1996; Kyle, Abell, & Roth, 1992). As Lederman (1992, p. 1012) reminded us: "We must *let the research questions direct the research approaches and data analysis procedures*" (original emphasis). If quantitative results cause problems of interpretation to practitioners (and others), we must find ways of making them more meaningful and more relevant. Often it is difficult to avoid statistical terminology if methods and results are to be reported in an unambiguous way. But there is something we can do easily, and that is to translate the results into something with which teachers are familiar; in other words, we need to go beyond the tests of statistical significance and report in terms of practical or educational value. I argue that such reporting is not only helpful to readers, but researchers are obliged to do it from a statistical point of view. Let us begin to examine this issue with some commentary from the *Journal of Research in Science Teaching.*

Plucker and Ball (1996) drew attention to the reporting of results in an earlier paper by Lazarowitz, Hertz-Lazarowitz, and Baird (1994), suggesting that particular findings described as significant in fact lacked practical significance. They pointed out that this could and should have been demonstrated using effect sizes, such as eta-squared. The authors (Lazarowitz et al., 1996, p. 682) replied that their findings were "of theoretical interest" even if the effect sizes (which they referred to as "this new procedure") were probably small.

Fundamentally, these comments highlight the old but important debate about the difference between statistical significance and the practical or educational value of research findings. The comments also highlight the contribution the notion of effect sizes can make to interpreting and reporting results, particularly as this contribution is far from new. Kirk (1996) attributed the idea of expressing the magnitude of an effect to Karl Pearson in 1901, and in 1925 Ronald Fisher proposed the use of eta (or the correlation ratio) as a measure of the strength of association between the independent and dependent variable. This procedure was recognized in statistics method textbooks during the 1930s (Huberty, 1993). Kirk noted that in 1969, Cohen's *d* (a standardized mean difference) became the first measure of effect magnitude which was explicitly called an effect size. During the 1970s, as the procedures of meta-analysis were developed and documented, measures of effect magnitude were thoroughly described (see, for example, Glass, 1977, and any number of textbooks on meta-analysis). The term "effect size" came into generic use, but I will use the term "effect magnitude" herein because it is more inclusive. Thus, we see that effect magnitudes and their use have been documented extensively for at least 2 decades; in fact, Glass published a descriptive paper on meta-analysis in the *Journal of Research in Science Teaching* in 1982 and the May 1983 issue of the *Journal* was dedicated to a comprehensive NSF-funded meta-analysis of research in science teaching from the 1950s through the early 1980s (see Anderson, Kahl, Glass, & Smith, 1983).

If effect magnitudes are not really new, are they important enough for researchers to chide each other about their use? The answer takes us back to the issue of the statistical and practical significance of the findings of quantitative research. There has been a groundswell of comment on two fronts. First, because statistical significance does not imply practical significance, researchers are urged (but apparently are reluctant) to address the issue of practical significance in reporting their results. The discussion about effect magnitude has sprung from this. The magnitude of the effect provides a quantitative estimate of practical significance, although the researcher is still required to interpret the educational importance of the result in the context of the research situation. Second, there is deep-rooted dissatisfaction with the theory and practice of the process of statistical significance testing (SST) itself. This dissatisfaction has resulted in a special issue of the *Journal of Experimental Education* (Thompson, 1993), strong statements in reworked editorial guidelines, such as those for *Educational and Psychological Measurement* (Thompson, 1994a) and the American Psychological Association (APA, 1995; Thompson, 1996,

1997), and recommendations for editors and others (Carver, 1993; Daniel, 1997; Shaver, 1993; Thompson, 1996) about the use of SST in papers to be published.

What is the concern all about, and should we, as researchers in science education, take heed? I will address three questions:

1. What are the problems associated with the use of SST?
2. How well do we as science education researchers report and discuss the statistical and practical significance of our findings?
3. How can we improve our own practice?

Problems Associated with the Use of SST

Most of the problems with SST stem from misunderstanding of what SST is about, with subsequent misinterpretation and misuse of the outcomes, but there is also disquiet about its theoretical underpinnings. Criticisms along both these lines have been published for more than 40 years (Cohen, 1994), often with the same authors reiterating their concerns because, apparently, they were unheard (see, for example, Carver, 1978, 1993). It is beyond the scope of this editorial to discuss fully the logic and criticisms of the process of SST, so I refer the reader to excellent discussions provided by Carver (1978), Cohen (1994), and Shaver (1993), among others. However, I offer an overview of some of the issues to provide context for the rest of my comments.

What Statistical Significance Testing Is About

Contrary to the impression given in many statistics method textbooks, SST is not a single, unified, unproblematic procedure. Huberty (1987) explained it as a hybrid of the philosophies of Fisher and of Neyman and Pearson. Fisher's significance testing procedure requires the statement of a null hypothesis based on a specified test statistic. Data are collected, the statistic is calculated, and the probability of the result is determined, given that H_O is true. H_O is rejected if p is small. The Neyman–Pearson hypothesis-testing approach requires the statement of both a null and an alternative hypothesis, together with a fixed-value probability (alpha) of rejecting the null hypothesis when it is true. Data are collected, the statistic is calculated, and if its value is in the rejection region, H_O is rejected in favor of H_1. Huberty's (1993) review revealed that early textbooks followed the Fisher approach and the Neyman–Pearson method, which introduced formally the alternative hypothesis and the idea of a critical region was incorporated into textbooks by 1950. Most current textbooks present a combination (usually unexplained) of these two approaches.

Shaver (1993) made a valuable contribution to the debate by explaining what SST is and what it is not. He provided a useful description of the meaning of SST:

> Our commonly used tests of statistical significance (z ratios, *t* ratios, and *F* ratios, such as in the analysis of variance or covariance) are procedures for determining the probability (usually at a pre-specified level of alpha) of a particular result, assuming the null hypothesis to be true, given randomization and a sample size of *n.* (p. 294)

There are three rather obvious and important aspects of this statement, each of which is fundamental to understanding what the process of SST is about. Each also is the subject of misinterpretation. These aspects are the nature of the probability that is determined in SST, the issue of randomization, and the implications of sample size.

Probability and the Null Hypothesis. The statement of probability in SST refers to the likelihood of a particular result from the data (D) given that the null hypothesis (H_O) is true; that is, $p(D|H_O)$. It is *not* a statement of probability about whether the null hypothesis is true given the data: that is, $p(H_O|D)$. Cohen (1994), Kirk (1996), and Menon (1993a), among others, pointed out that there is a widespread and erroneous belief that a low probability for the first statement implies a low probability for the second—that is, the false assumption that $p(D|H_O) = p(H_O|D)$. Carver (1978, p. 384) demonstrated the error of this assumption in a dramatic way:

> What is the probability of obtaining a dead person (label this part D) given that the person was hanged (label this part H); that is, in symbol form, what is $p(D|H)$? Obviously it will be very high, perhaps .97 or higher. Now let us reverse the question. What is the probability that the person was hanged (H) given that the person is dead; that is, what is $p(H|D)$? This time the probability will undoubtedly be very low, perhaps .01 or lower.

Carver (1978, p. 385) pointed out that although substituting the first estimate (.97) for the second (.01) "seems to be an unlikely mistake, it is exactly the kind of mistake that is made with interpretations of statistical significance testing—by analogy, calculated estimates of $p(H|D)$ are interpreted as if they were estimates of $p(D|H)$." Unfortunately, it is easy to find examples of these mistakes, often in textbooks, as Cohen (1994) and others have pointed out.

Randomization and Replicability. Clearly, because SST assumes that H_O is true, it makes sense only if the data come from samples randomly selected from the (often hypothetical) population to which H_O refers. This enables statements of probability to be made about the likelihood that the sample came from this population. Misinterpretations commonly arise. The first is the misunderstanding that a random sample is always representative of its population. However, random sampling does not ensure representativeness in a single sample; only repeated random sampling can ensure representativeness in the long run. Similarly, random assignment can meet only the assumption of nonsystematic between-group differences; it does not ensure that groups are equivalent in a single study. Thus, the result cannot be used to make a statement about the results of another study. Carver (1978) and Menon (1993a) both quoted samples where authors misinterpreted a statistically significant finding to mean that the result will replicate in a new study. Just as the result of one coin toss can have no bearing on the result of a second, a statistically significant result for one sample provides no information about the results from the next sample; it does not imply that the result will replicate in the next study.

Another misinterpretation is that a statistically significant result can be interpreted in a causal way; that the rejection of the null hypothesis confirms the theory underlying the research hypothesis. A statistically significant result does not indicate that the treatment is responsible for any effect. Again, examples of this misinterpretation are easy to find.

All of this underscores the importance of replication to have confidence in the outcomes of research. If the study is replicated with other samples, then consistent findings will provide confidence in the result. Even if the differences are small and nonsignificant in a statistical sense, if they replicate, then it is more likely that they are real.

Size of Sample. The size of the sample identifies the theoretical distribution of the statistic being tested, and every statistic has its degrees of freedom based on *n*. Thus, the calculated probability is a function of the sample size, as well as any effect which exists. The larger the sam-

ple, the smaller the sampling error and the more likely that the calculated statistic will be statistically significant. The advantage of larger samples is that they are more likely than small ones to resemble the population in the long run, so a larger sample increases the power of the statistical test, making it more likely to detect a difference if one exists.

An important point to note here is that a statistically significant result does not mean that the effect is large or that the finding is necessarily important. Trivial effects may be statistically significant if the sample is large enough. Of course, the reverse is also true: nonsignificant findings may also be important. It is the responsibility of the researcher to address the issue of importance in the context of the research.

Theoretical Problems

Misinterpretation and misuse of SST can be prevented by better understanding of what it is about, but even when SST is flawlessly used and interpreted, there are still aspects of fundamental concern to its critics. The first problem is that SST does not tell us what we want to know. As Cohen (1994) and Kirk (1996) pointed out, the question of interest to the researcher is: What is the probability that the null hypothesis is true, given that I have these data? This is $p(H_O|D)$, but SST considers $p(D|H_O)$, which, as we have seen, are not the same. SST proides a value of p which, if it is considered small enough, suggests that chance is an unlikely explanation for the data, and the null hypothesis is rejected. Instead, the emphasis should be on the data and whether they support the research hypothesis. We should note that Bayes' theory and Bayesian statistics do address the issue of $p(H_O)$, or the probability of the null hypothesis before the experiment begins (Cohen, 1994), and have much to offer the debate on statistical inference (Edwards, Lindman, & Savage, 1963). However, at this time Bayesian methods are rarely used in science education research, and until they have evolved it seems more important to understand, and thus better use, the techniques we presently employ.

Other theoretical problems relate to the assumptions which underlie SST. In the first place, random sampling is seldom possible in educational research; random assignment of a treatment or intervention is more common but still unusual. Since randomization is essential for SST (Shaver, 1993), the violation of the assumption of randomness makes a mockery of the test. In the second place, as most critics point out, the null hypothesis is never exactly true. It is therefore a trivial exercise to attempt to reject it. Cohen (1994) added the point that the almost universal assumption that H_O is equivalent to an effect size of zero is "downright ridiculous" (p. 1000).

In view of these problems with SST, why do researchers cling to this procedure? Why have researchers adhered to the notion of statistical significance instead of looking for and making statements about the practical or educational significance of their findings? One reason often postulated is the seemingly objective nature of SST. By finding a value for p and using it to accept or reject H_O, a neat, dichotomous, and deceptively simple decision is delivered into the hands of the potentially subjective researcher. In fact, what SST does is divert attention from the research hypothesis. Carver (1978) compared two kinds of research methods. The scientific method moves directly from data collection to the question: Do the data support the research hypothesis? If they do, then alternative hypotheses are considered, including H_O. In the other method, which Carver called corrupt, the research process moves from data collection to a test of the statistical significance of the data, invoking H_O. In this method, "a result that is not statistically significant is automatically interpreted as providing no support for the research hypothesis, no matter how much the data tend to confirm it" (Carver, 1978, p. 390).

Should SST Be Abandoned?

Many critics of SST believe that its continued use impedes the growth of cumulative knowledge (Cohen, 1994; Schmidt, 1996) and that it should be abandoned (Carver, 1978, 1993; Cohen, 1994; Menon, 1993a) or at least minimized (Shaver, 1993). Others believe that the problems lie not with SST but with its misinterpretation. Huberty (1993) thought that some of the blame for this lies with poor presentation by textbooks, teaching and reporting, and poor journal editorial review. More moderate commentators, including some journal editors, suggested that, when properly used, SST still has something to offer to the educational researcher (Asher, 1993; Bourke, 1993; Clements, 1993; Levin, 1993; Robinson & Levin, 1997; Rowley, 1993; Schafer, 1993; Zwick, 1997).

There is widespread agreement among authors that evidence beyond SST should be presented in the reporting of research. One universal recommendation is for the use of estimates of the size of the effect, generically called "effect sizes," but also termed "effect magnitudes" (Kirk, 1996) or "magnitude-of-effect estimates" (Snyder & Lawson, 1993). These tend to be measures of the strength of association between the variables, such as a correlation coefficient and its square, or a standardized mean difference, but there are also others (Kirk, 1996; Snyder & Lawson, 1993; Tatsuoka, 1993). Of course, effect magnitudes are not unproblematic. Because least-squares methods of regression analyses capitalize on sampling error, correlational measures of effect magnitude tend to be biased upward, and a range of procedures to correct the bias have been suggested (Snyder & Lawson, 1993; Thompson, 1993). In terms of this discussion, a major advantage is that effect magnitude refocuses attention on the data and takes it off the null hypothesis (Kirk, 1996). Another, of course, is that effect magnitudes highlight the distinction between statistical and practical significance (Snyder & Lawson, 1993). As Tatsuoka (1993, p. 461) pointed out, the rejection of H_O "does not suggest that the effect is large or even nontrivial but simply that it is non zero. Hence the rejection of H_O 'cries out' for estimating the magnitude of the effect in question."

Reporting of Statistical and Practical Significance in Science Education Research

Plucker and Ball's (1996) comment in the *Journal of Research in Science Teaching* about the benefits of reporting effect magnitudes as a way of highlighting practical significance was a timely one. To obtain a broad perception of how common is the use of effect magnitude in interpretation and reporting of science education research, I surveyed the 1996 volumes of five English language science education journals: *Journal of Research in Science Teaching* (JRST), *International Journal of Science Education* (IJSE), *Research in Science Education* (RISE), *Research in Science and Technological Education* (RSTE), and *Science Education* (SE). Each issue was scanned, the kind of the research reported in each article was considered (comments and editorials were excluded), and the nature of the data analysis and reporting were noted. The results are presented in Table 1.

Altogether, a total of 197 articles were scanned, and 119, or 60%, of these reported some kind of numerical data. More than half of these articles (56%) reported frequency data in tables or graphs, often to represent the number of students with particular views or ideas, or who correctly completed a task. The remaining 52 articles, representing about one quarter of all of those published, could be described as using quantitative methods and reported statistical analyses of some kind. Nine of these 52 articles did not deal with effects, so the reporting of effect magnitudes was not relevant. The other 43 dealt with inferential statistics where effect magnitudes

Table 1
Numbers of papers reporting effect magnitudes in five science education journals

Variable	*JRST*	*IJSE*	*RISE*	*RSTE*	*SE*	Total	% of Total
Total papers published	52	67	32	16	30	197	100
No numerical data reported	20	31	10	1	16	78	40
Frequency data only reported	16	24	12	8	7	67	34
Statistical analysis reported	16	12	10	7	7	52	26
Effect magnitudes not relevant	4	1	2	1	1	9	5
Effect magnitudes are relevant	12	11	8	6	6	43	22
Effect magnitudes are reported	2	2	2	2	2	10	5

could have been calculated and reported to clarify the findings. Only 10 of these 43 articles, less than a quarter, reported any effect magnitudes. Thus, we find that effect magnitudes appeared in 5% of the published articles in the five journals, or, more important, only 22% of those articles where effect magnitudes were relevant actually included them. Furthermore, in several of these articles, the effect magnitudes were reported in tables but were not interpreted in the text.

A comparison can be made between the reporting of quantitative research involving inferential statistics in science education and in the psychological literature. Kirk (1996) reviewed the contents of the 1995 volumes of four journals: *Journal of Applied Psychology, Journal of Educational Psychology, Journal of Experimental Psychology, Learning & Memory,* and the *Journal of Personality and Social Psychology.* The numbers of articles using inferential statistics in each journal were, respectively, 57, 49, 111, and 174, and, again, respectively, the percentages of these which included at least one measure of effect magnitude were 77%, 55%, 12%, and 46%. This compares with 22% for the science education journals reported in Table 1.

The most frequently used inferential procedures in the psychological journals were analyses of variance, the *t* test for mean differences, and regression analysis. Based on Table 5 in Kirk (1996), the most common kind of effect magnitude reported was variance-accounted-for (in around 90% of papers). Only 4% were standardized mean difference. The pattern was similar in the 1996 science education journals. The common inferential procedures involved analyses of variance, correlation/regression, chi-square, and the *t* test. The small number of effect magnitudes reported were mainly squared correlation coefficients, although they were not always interpreted as measures of the magnitude of the effects.

In sum, this survey suggests that of the substantial number of research papers published in the science education literature which employ inferential statistics, a minority report effect magnitudes and even fewer provide discussion of, or even the distinction between, statistical and practical significance of the findings. Based on these two surveys, the performance of our journals appears to lag behind those in the psychological literature. It is not clear why this is so, but a possible explanation is that psychology journals give more attention to issues relating to SST. In this context, we might note that Schmidt's (1996) oft-quoted paper in which he argues that reliance on SST has "systematically retarded the growth of cumulative knowledge in society" (p. 115) was based on his presidential address to the Division of Evaluation, Measurement, and Statistics of the American Psychological Association. Whatever the reason for the difference, it seems there is considerable room for improving our practice in reporting science education research.

Improving the Quality of Reported Research in Science Education

Reference has already been made to a number of papers devoted to discussion about significance in educational and psychological research. Many suggest guidelines for improved reporting of quantitative research, including Carver (1978, 1993), Cohen (1994), Daniel (1997), Shaver (1993), Thompson (1993, 1994a, 1996), and Thompson and Snyder (1997). These and other discussions have been considered in drawing up a list of suggestions for authors in science education research. It might be worth noting that these ideas are important in planning the research as well; they are not simply afterthoughts!

Use Correct Terminology

A simple contribution that authors can make to minimize the emphasis placed on SST in the reporting of research findings is to insert the word "statistically" before significant. Only a couple of articles in the science education journals surveyed followed this practice, and then not consistently. As Carver (1993) pointed out, many readers do not know that significant almost always means statistically significant. Menon (1993b) argued that points like this are worth saying again, because misunderstanding and misinterpretation of SST are still widespread. We need to check our own reports to ensure that we do not unconsciously contribute to further confusion. I recommend that you review Shaver's (1985a, 1985b) two conversational pieces on interpreting statistical tests of significance.

Provide Sufficient Information about the Data

Researchers need to keep the research data in focus by describing them thoroughly. Cohen (1994) talked about exploratory data analysis as a way of understanding the data. In any case, sufficient information should be provided to readers to allow them to make some judgments of their own about the findings. At least point estimates and an indication of sampling error should be reported. For example, means accompanied by standard deviations or standard errors can provide this information. Line graphs are clear and easy to inspect, but they do not routinely indicate sampling error unless error bars are added. In this vein, there has been considerable attention paid to the reporting of confidence limits, instead of point estimates. Kirk (1996, p. 745) noted that

> A confidence interval contains all of the information provided by a significance test and, in addition, provides a range of values within which the true difference is likely to lie . . . and makes trivial effects harder to ignore.

Serlin (1993) explained the idea of using range null hypotheses based on confidence limits instead of point null hypotheses. However, confidence intervals are based on the logic of statistical significance testing so their use does not solve all of the problems of SST (Zwick, 1997). Furthermore, confidence intervals for many statistics are not well understood (Behrens, 1997; Zumbo, Pope, & Stork, 1997). At the moment, they are not widely reported in educational research. Dugan, Huston, Franz, and Izzo (1997) examined 681 articles and found only eight that reported confidence intervals, with most leaving them uninterpreted.

Another technique which can be used by the researcher to aid interpretation of the findings is to provide additional information based on theoretical, rather than empirical, extrapolation of the results to different sample sizes. This is sometimes called what-if analysis. For example, Thompson (1994a) suggested that authors might explicitly index the statistical significance of

their statistic to the sample size; however, this approach is not without criticism (Robinson & Levin, 1997).

Replication

Replication remains the fundamental method by which confidence is established in the findings of research using inferential statistics. Conducting another study with a new sample is the best way, but in the real world of schools and teachers whose priorities often do not match those of the researcher, this kind of external replication is rarely possible. A number of internal methods of replication which use a single data set are in use. The first of these is cross-validation, where the data are divided and the analyses replicated in each part. Other empirical methods include bootstrap and jack-knife techniques which attempt to estimate the likelihood that results will replicate. These methods are well described by Thompson (1994b), but although helpful, they are not as persuasive as using new data.

Report and Interpret Effect Magnitudes

Ideally, we would like the effects of our educational experiments to be dramatic and indisputable. Unfortunately, they rarely are. Not many of our results pass what Edwards et al. (1963, p. 217) described as the interocular traumatic test: that is, "You know what the data mean when the conclusion hits you between the eyes." Usually, we have to rely on some other way to convey the magnitude of our findings.

The use of estimates of the effect magnitude as a complement to SST is a main theme of this editorial. In the surveys noted above, the most frequently reported measure of effect magnitude (although not always interpreted) is a squared correlation coefficient, R^2, r^2, or eta-squared. These measures give the proportion of variation in the scores of the dependent or criterion variable (such as scores on a science achievement test) which can be predicted from the variance of the independent or predictor variable(s) (such as the kind of teaching approach used). Thus, if $r^2 = 0.5$, half or 50% of the variance in one variable is shared with the other. In terms of effect magnitude, Kirk (1996) suggested this might be described as a large effect. Whether this result is statistically significant depends on the size of the sample, but given the context and circumstances of the research, the researcher might decide, for example, that only effect magnitudes $> r^2 = 0.1$ are important enough to merit concern.

The second most frequently reported kind of effect magnitude is a standardized mean difference. This is the measure most commonly used in meta-analysis because it translates the effects found in different studies to measures on a similar metric. Several researchers estimate that an effect magnitude of about 0.5 is medium (Kirk, 1996). With large groups, an effect magnitude of only 0.1, one tenth of a standard deviation, may be statistically significant, even though it may not be noticeable. Depending on the circumstances and the context, the researcher might consider that an effect size of 0.2 is needed to indicate a worthwhile effect.

Notice a subtle shift in emphasis here: The interpretation of the effect magnitude as a worthwhile effect places the onus on the researcher to decide whether the result has practical or educational value, an important and worthwhile step beyond the reporting of statistical significance. It is incumbent upon researchers to report and interpret effect magnitudes, whether or not they are significant. Further useful discussion of effect magnitudes is provided by Tatsuoka (1993) and Kirk (1996), their use in repeated measures designs is documented by Dunlap, Cortina, Vaslow, and Burke (1996), and the process of correction for bias is described by Snyder and Lawson (1993).

Summary

I began with the suggestion that improvement in the interpretation and reporting of quantitative research findings would help bridge the gap between research and practice in science education. Attention has been drawn to several misinterpretations of SST: namely, that it is a test of the null hypothesis (rather than the data), that a statistically significant finding provides evidence that the result is replicable, that the effect is caused by the treatment, and that it is large and/or important. All of these misinterpretations can be found in science education literature, but probably little purpose would be served by drawing attention to specific examples.

A survey of recent published articles in science education revealed that considerable scope exists for improvement in conveying and interpreting the findings of quantitative research. Some suggestions have been made to enhance the quality of reported research. The bottom line is that researchers need to address the importance of their findings, giving consideration to factors such as the impact or benefit and the cost of producing it, in terms of time, effort, money, and human values. Research interpreted only in terms of the statistical significance of its findings is not only barren but unfriendly to the practitioners whose reward for allowing us to research in their classrooms should, at least, be a clear statement of the meaning, importance, and relevance of the results.

An earlier version of this article was presented at the annual conference of the Australasian Science Education Research Association, Adelaide, July 1997. The author thanks Ric Lowe and Russell Jones for helpful comments on the earlier draft, and Dick Gunstone for drawing the author's attention to the interocular traumatic test as described by W. Edwards, H. Lindman, and L.J. Savage in 1963.

References

Amabile, T.M. (1982). Conversation I: The gap between teachers and researchers. In T.M. Amabile & M.L. Stubbs (Eds.), *Psychological research in the classroom: Issues for educators and reseachers* (pp. 9–20). New York: Pergamon Press.

Amabile, T.M., & Stubbs, M.L. (1982). *Psychological research in the classroom: Issues for educators and researchers.* New York: Pergamon Press.

American Psychological Association. (1995). *Publication manual of the American Psychological Association* (4th ed.). Washington DC: Author.

Anderson, R.D., Kahl, S.R., Glass, G.V., & Smith, M.L. (1983). Science education: A meta-analysis of major questions. *Journal of Research in Science Teaching, 20,* 379–385.

Asher, W. (1993). The role of statistics in research. *The Journal of Experimental Education, 61,* 388–392.

Behrens, J. (1997, March). *The logical and historical bases and inferential limitations of confidence intervals.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Bourke, S. (1993). Babies, bath water and straw persons: A response to Menon. *Mathematics Education Research Journal, 5,* 19–22.

Caliendo, S.M., & Kyle, W.C., Jr. (1996). Establishing the theoretical frame. *Journal of Research in Science Teaching, 33,* 225–227.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Carver, R.P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education, 61,* 287–292.

Clements, M.A. (1993). Statistical significance testing: Providing historical perspective for Menon's paper. *Mathematics Education Research Journal, 5,* 23–27.

Cohen, J. (1994). The Earth is round ($p<.05$). *American Psychologist, 49,* 997–1003.

Daniel, L.G. (1997, March). *Statistical significance testing in* Educational and Psychological Measurement *and other journals.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Dugan, J., Huston, D., Franz, S., & Izzo, R. (1997, March). *How confidence intervals are used in educational research.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Dunlap, W.P., Cortina, J.M., Vaslow, J.B., & Burke, M.J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170–177.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Glass, G.V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5,* 351–379.

Glass, G.V. (1982). Meta-analysis: An approach to the synthesis of research results. *Journal of Research in Science Teaching, 19,* 93–112.

Huberty, C.J. (1987). On statistical testing. *Educational Researcher, 16,* 4–9.

Huberty, C.J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman–Pearson views in textbooks. *The Journal of Experimental Education, 61,* 317–333.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Koballa, T., Jr. (1997). Two communities: One challenge. Address from exiting President of NARST. *NARST News, 40,* 3–4.

Krockover, G.H., & Shepardson, D.P. (1995). Editorial: The missing links in gender equity research. *Journal of Research in Science Teaching, 32,* 223–224.

Kyle, W.C., Jr., Abell, S.K., & Roth, W.-M. (1992). Toward a mature discipline of science education. *Journal of Research in Science Teaching, 29,* 1015–1018.

Lazarowitz, R., Hertz-Lazarowitz, R., & Baird, J.H. (1994). Learning science in a cooperative setting: Academic achievement and affective outcomes. *Journal of Research in Science Teaching, 31,* 1121–1131.

Lazarowitz, R., Hertz-Lazarowitz, R., & Baird, J.H. (1996). Reply: Learning and teaching science in the classroom: New settings and statistical comments. *Journal of Research in Science Teaching, 33,* 681–683.

Lederman, N.G. (1992). You can't do it by arithmetic, you have to do it by algebra! *Journal of Research in Science Teaching, 29,* 1011–1014.

Levin, J.R. (1993). Statistical significance testing from three perspectives. *The Journal of Experimental Education, 61,* 378–382.

Menon, R. (1993a). Statistical significance testing should be discontinued in mathematics education research. *Mathematics Education Research Journal, 5,* 4–18.

Menon, R. (1993b). Take off those blinkers, mate! Response to Bourke, Clements and Rowley. *Mathematics Education Research Journal, 5,* 30–33.

Pekarek, R., Krockover, G.H., & Shepardson, D.P. (1996). The research–practice gap in science education. *Journal of Research in Science Teaching, 33,* 111–113.

Plucker, J.A., & Ball, D. (1996). Comment on "Learning science in a cooperative setting: Academic achievement and affective outcomes." *Journal of Research in Science Teaching, 33,* 677–679.

Robinson, D.H., & Levin, J.R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26,* 21–26.

Rowley, G. (1993). Response to Menon. *Mathematics Education Research Journal, 5,* 28–29.

Schafer, W.D. (1993). Interpreting statistical significance and nonsignificance. *The Journal of Experimental Education, 61,* 383–387.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 15–129.

Serlin, R.C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *The Journal of Experimental Education, 61,* 350–360.

Shaver, J.P. (1985a). Chance and nonsense: A conversation about interpreting tests of significance: Part 1. *Phi Delta Kappan, 67,* 57–60.

Shaver, J.P. (1985b). Chance and nonsense: A conversation about interpreting tests of significance: Part 2. *Phi Delta Kappan, 67,* 138–141. Erratum, 1986, *67,* 624.

Shaver, J.P. (1993). What statistical significance testing is, and what it is not. *The Journal of Experimental Education, 61,* 293–316.

Shymansky, J.A., & Kyle, W.C., Jr. (1992). Establishing a research agenda: Critical issues of science curriculum reform. *Journal of Research in Science Teaching, 29,* 749–778.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education, 61,* 334–349.

Stubbs, M.L. (1982). Conversation II: Issues in the application of research results. In T.M. Amabile & M.L. Stubbs (Eds.), *Psychological research in the classroom: Issues for educators and researchers* (pp. 21–35). New York: Pergamon Press.

Tatsuoka, M. (1993). Effect size. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 461–479). Hillsdale, NJ: Erlbaum.

Thompson, B. (1994a). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837–847.

Thompson, B. (1994b). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality, 62,* 157–175.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25,* 26–30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26,* 29–32.

Thompson, B. (Ed.). (1993). Statistical significance testing in contemporary practice: Some proposed alternatives with comments from Journal editors [Special issue]. *Journal of Experimental Education, 61*(4).

Thompson, B., & Snyder, P.A. (1997, March). *Use of statistical significance tests and reliability analyses in published counseling research.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

White, R.T. (1984). Research, and the end of schools as we know them. *Australian Journal of Education, 28,* 3–16.

Zumbo, B., Pope, G.A., & Stork, J. (1997, March). *Are confidence intervals useful? Looking beyond the simple cases.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Zwick, R. (1997, March). *Would the abolishment of significance testing lead to better science?* Paper presented at the annual meeting of the American Educational Research Association, Chicago.