Mathematics Education Research Journal 1993, Vol. 5, No. 1, 4–18

Statistical Significance Testing Should be Discontinued in Mathematics Education Research

Rama Menon
Nanyang Technological University

It is claimed here that the confidence mathematics education researchers have in statistical significance testing (SST) as an inference tool par excellence for experimental research is misplaced. Five common myths about SST are discussed, namely that SST: (a) is a controversy-free, recipe-like method to allow decision making; (b) answers the question whether there is a low probability that the research results were due to chance; (c) logic parallels the logic of mathematical proof by contradiction; (d) addresses the reliability/replicability question; and (e) is a necessary but not sufficient condition for the credibility of results. It is argued that SST's contribution to educational research in general, and mathematics education research in particular, is not beneficial, and that SST should be discontinued as a tool for such research. Some alternatives to SST are suggested, and a call is made for mathematics education researchers to take the lead in using these alternatives.

Because statistical significance testing (SST) makes use of mathematics, many researchers assume that SST is supported by mathematical logic. So it is not surprising that inferences based on SST are used to strengthen the credibility of mathematics education research (MER). For example, although about 43% of articles in the last 10 issues (1990-1991) of the *Journal of Research in Mathematics Education (JRME)* reported quantitative research, all of these reported significance levels, with none making any reference to other statistical measures, such as effect size.

In this paper I intend to show that the confidence MER seems to have in SST is misplaced, by first elucidating five common myths about SST. These myths are that SST: (a) is a controversy-free, single, unified, recipe-like method to allow decision making in the face of uncertainty; (b) answers the pressing question whether there is a low probability that the research results were due to chance; (c) logic parallels the logic of mathematical proof by contradiction; (d) addresses the reliability/ replicability question; and (e) is a necessary but not sufficient condition for the credibility of results.

Then, I suggest that SST's contribution to educational research in general, and MER in particular, is both erroneous and harmful, and that SST should be discontinued as a tool for such research. Next, I list, without elaboration, a few suggested alternatives to SST, referring the reader to more comprehensive treatments of these alternatives by other writers. I will conclude by stating that these suggested alternatives are only partial solutions to the question of replicability and generalisability and that the only appropriate solution is actually to replicate the study a number of times.

More than 60 years ago, Tyler (1931) cautioned against the uncritical use of statistical significance testing, warning researchers that a statistically significant difference was not necessarily an important difference, and a difference that was not statistically significant may be an important difference. Since then, there have been many critics of SST (e.g., Carver, 1978; Coats, 1970; Cronbach & Snow, 1977; Dar, 1987; Falk, 1986; Falk & Greenbaum, 1993; Guttman, 1977, 1985; Morrison & Henkel, 1969; Rosnow & Rosenthal, 1989; Shaver, 1992). Repeated calls for caution when making inferences from SST (e.g., Bakan, 1966; Carver, 1978; Coats, 1970; Cohen, 1977; Cooper, 1984; Cronbach & Snow, 1977; Falk, 1986; Falk and Greenbaum, 1993; Hays, 1974; Levy, 1967; Morrison & Henkel, 1969; Pauker & Pauker, 1979; Shaver, 1985a,1985b; Slakter, Yu, & Suzuki-Slakter, 1991; Stevens, 1968; Thompson, 1992; Tyler, 1931), have had limited effect on many education researchers who have continued to use experimental designs.

More recently, in a symposium at the annual meeting of the American Educational Research Association, Shaver (1992) stated that "a quick perusal of educational research journals, educational and psychological statistics textbooks, and doctoral dissertations will confirm that tests of statistical significance continue to dominate the interpretation of quantitative data in educational research" (p. 1).

MYTH # 1:
A Controversy Free, Single, Unified Approach to Decision Making

Although most people know of Fisher's contribution to statistical theory, few are aware that many other equally respected statisticians (and psychologists) did not agree with his theories. For example, Neyman believed that Fisher's methods of testing were, in a "mathematically specifiable sense, worse than useless" (Stegmuller, 1973, p. 2, cited in Gigerenzer & Murray, p. 17). As well, Meehl (1978, p. 817) said, "Fisher has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis ... is ... one of the worst things that ever happened in ... psychology." And, according to Gigerenzer and Murray (1987), "You won't catch Jean Piaget or Wolfgang Kohler calculating a *t* or *F* value, and ... eminent figures ... Bartlett, Stevens and Skinner all explicitly rejected statistical inference. They preferred to trust their own informed judgment" (p. 26).

What were some of the controversies? Firstly, there were many differences between the Neyman-Pearson and the Fisher theories of SST. The former introduced the idea of *symmetric* hypothesis testing, involving both H_0 and H_1, in contrast to Fisher's *asymmetric* hypothesis testing involving only H_0. They introduced Type I and Type II errors, and were emphatic that rejection or acceptance of H_0 or H_1 did not imply belief in or disproof of either hypothesis. For them, hypothesis testing did not give a single best method of inference: effect size and a subjective cost benefit analysis (balancing losses against gains), had also to be taken into account, not just the significance level, and all of these would vary from context to context. Unlike Fisher, they made a distinction between the *mathematical* and *subjective* parts of SST and decision making, saying that the subjective parts were those not directly quantifiable, but were qualitative factors such as benefits or drawbacks that might result. Moreover, their statistical decision theory was restricted to applications where repeated random sampling from a defined population was a reality (so the theory cannot apply to MER in general, as very seldom does MER, especially involving intact classes, make use of random samples of students, taken repeatedly). That is, they disagreed with Fisher's idea of non-random sampling from an undefined, hypothetical population.

Kendall (1943), the originator of Kendall's tau, also considered Fisher's notion of a random sample from a hypothetical infinite population "a most baffling conception" (p. 17). This notion is certainly baffling, given that Fisher's experiments were all on small, non-randomly selected samples, with undefined populations.

Lately, others have stressed the importance of randomisation for SST. For example, Shaver (1992) said "randomization is *essential* to the typical tests of statistical significance" (p. 4), and Glass and Hopkins (1984) pointed out that "inferential statistics is based on the assumption of random sampling from populations" (p. 177). As well, Winer, Brown, and Michels (1991), in discussing ANOVA assumptions, categorically stated:

> Violating the assumption of random sampling of elements from a population and random assignment of the elements to the treatments may totally invalidate any study, since randomness provides the assurance that errors are independently distributed within and between treatment conditions and is also the mechanism by which bias is removed from treatment conditions. (p. 101)

Admittedly, there are some who believe that randomisation is not essential for SST. For example, Thompson (1987) claimed that "significance testing imposes a restriction that samples must be representative of a population, but does not mandate that this end must be realized through random sampling" (pp. 8–9). But, as Shaver (1992) pointed out, randomness and representativeness are related only to the extent that there is repeated random sampling to ensure representativeness in the long run, not that a single random sample is wholly representative of the population, only that sample characteristics may differ by chance from the population characteristics.

Moreover, only where repeated random sampling is a reality (as in say, quality control experiments) can one talk meaningfully about the level of significance (as the long run frequency of Type I errors). In other words, the significance level is meaningless for a *single* experiment as the inductive logic behind SST demands replication.

Fisher (1960) did say that "no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon. ... In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure" (pp. 13–14). His argument implied that replication is necessary, especially as he was using an inductive argument for inferences drawn from SST.

However, such an argument led to an inconsistency in Fisher's logic, referred to as *The Experimental Paradox* by Spencer-Brown (1957, pp. 64–66). In essence, Fisher (1960) suggested that a *series* of replications be regarded as a *single* experiment. This was because the logic of Fisherian SST depended crucially on "its being unique" (Falk & Greenbaum, 1993, p. 13), or the *once-ness* of the experiment (Bakan, 1966). The problem was to reconcile the necessity of uniqueness with that of replication—hence the paradox.

Prior to World War II, textbooks only presented Fisher's single instrument for inductive inference, namely that of disproving H_0. However, when Neyman-Pearson theories became better known, textbook writers came up with a "hybrid" theory, the result of retaining both Fisher's idea of disproving H_0, and Neyman-Pearson's consideration of Type II errors, even though Type II errors had "no meaning in the context of null hypothesis testing nor could it be determined" (Gigerenzer & Murray, 1987, p. 21).

Thus, the differences between the Fisher and Neyman-Pearson theories (as well as the substance of the numerous acrimonious debates generated by these differences) were completely ignored, and many researchers and later textbook writers slavishly followed a seemingly controversy-free, single, unified idea of SST. This state of affairs could be attributed to viewing the textbook not as "as a rational reconstruction of selected items of knowledge" but rather as "the body of achieved knowledge containing inviolable truths and predictions" (Factor & Kooser, 1981, p. 28).

Be that as it may, few researchers seem to be aware that, although both Fisher and Neyman-Pearson seemed to hold the frequentist perspective, they disagreed on basic issues such as randomisation, the need for an alternative hypothesis, and whether to make decisions on an "objective" reject-accept dichotomy or include "subjective" considerations as well. Mathematics education researchers, in particular, seem to be drawn to the neat categories of dichotomous decisions engenderd by the supposedly objective SST.

MYTH # 2:
SST Implies a Low Probability that Chance Affected the Result

When H_0, the hypothesis of "no difference" or "hypothesis of chance," is used, we can estimate the unknown variability of the population by using the known variability of the sample. Then, using the sample size and the estimated population variance, we can compute the frequency of the expected sampling error or mean difference (between population and samples) of any particular magnitude. Computations from, say, a *t* test will then give us a *p* value, such as $p = .01$.

This, in effect, means that only in 1 out of 100 pairs of samples could we expect a mean difference of, say, 5 or larger, if chance alone were acting (i.e. assuming that chance alone were working 100% of the time). That is, we would still be making an error 1% of the time by rejecting H_0, when H_0 were actually true. But we do not know that H_0 is really true.

Note that this *p* value has been calculated by first of all assuming that chance alone was working all the time (see Carver, 1978, where he discusses the "odds-against-chance fantasy"). So, how could we say that chance affected the results only 1% of the time, when we not only computed this *p* value of 1% by assuming that chance did cause the mean difference 100% of the time, but also used this *p* value to decide whether to accept or reject the notion that 100% of the time chance did cause the mean difference?

In other words, we do *not* know what proportion of samples would show *that* chance affected the results. For that we would need 100 pairs of samples partitioned into two groups where H_0 was true for one group of means and H_0 was not true for the other group of means. Then, if the expected frequency was 1 out of 100 when H_0 is true, then we could immediately compute that 99 of the 100 would occur when H_0 is not true.

So, statistical significance means statistical rareness, or a low probability of getting such results, when H_0 is assumed to be true; it tells us nothing about the probability that H_0 is true (i.e. that the probability is low that the results were due to chance), even though many of those using SST believe that they are answering the question "could these results have occurred very rarely (i.e. with a low probability) by chance?"

Although this myth has been discussed before (Dawes, 1981; Diaconis & Freedman, 1981; Falk, 1986), it is worth further elaboration. For example, let $P[D \mid H_0]$ mean the probability of a person dying, given that she has cancer, and let $P[H_0 \mid D]$ mean the probability a person has cancer, given that she is dying (Carver, 1978, gave a similar example). No one would mistake $P[D \mid H_0]$ for $P[H_0 \mid D]$, but that is exactly what is being done when one infers from a significance level of .05 that there is, at most, a .05 probability that chance "caused" the result.

This widespread fallacy is exemplified below, by a report on the effectiveness of writing-to-learn approach to learning mathematics (Lesnak, 1989):

> Using a statistical test of the difference between two means, this difference of 3.2% in their final averages is significant at a level of significance of 4.6%. Very briefly, this means that if all variables in the two groups were equal except for the use of writing-to-learn activities, then the probability of this difference occurring by Chance is less than 4.6%. (p. 154)

The researcher quoted above has had 25 years of teaching experience in mathematics and a good statistical background. He was studying students undertaking remedial algebra at a college and was actually trying to answer this question: I have got these results now. Given that I have these results, what is the probability that chance caused these results i.e., $P[H_0) \mid D]$? He had made the common error of inferring that the probability was .046 that, given the results, these results were due to chance, i.e. $P[H_0 \mid D] = .046$. Actually, the .046 referred to is $P[D \mid H_0]$, the probability of obtaining such results, given that chance alone was operating. In other words, he had already assumed that chance alone was working 100% of the time, before computing the .046 (see Carver, 1978, for an elaboration of this fallacy).

According to Falk (1986), psycholinguistic factors could explain, in part, the reason for the pervasiveness of this error. For example, let R be the event that a sample result is in the rejection region, and let H_0 be the event that chance alone is operating. Then, $P[R \mid H_0] = P[\text{Type I Error}] = \alpha$, the level of significance. However, the phrase "Type I Error" triggers off the notion of a simple event, probalistically-speaking. That is, one tends to forget that one is talking about a *conditional probability,* which necessarily involves at least two events (in this case R and H_0). This leads to the idea of an error, an error interpreted in terms of the conjunction of the two events R and H_0, in either one of the following ways:

1. H_0 is true *and* it is subsequently rejected; or
2. H_0 is rejected *and* subsequently found to be true.

That this linguistic confusion exists even among those critical of SST is clearly shown by the following statement by Levy (1967), cited in Falk (1986): "Statistical significance refers only to the reliability of an obtained result, the confidence with which a null hypothesis may be rejected or the probability that a Type I error has been committed" (p. 90).

As Falk pointed out, many misconceptions have "subtly crept in" (p. 90). For example, "the confidence with which a null hypothesis may be rejected" implies the degree of belief *after* the results have been obtained, not in the degree of belief in the results, given the hypothesis. The use of the present perfect tense ("has been committed") in the statement "the probability that a Type I error has been committed" shows that rejection of H_0 has *already* taken place, given the results—which is the inverse of the probability of getting the results, given H_0.

So, even though the level of significance only answers the question "what is the probability of the results occurring, *when* chance alone were operating?", one believes that one has answered the question "what is the probability *that* chance alone was operating, given the results?", which is further shortened to "could the results have occurred *by* chance?".

This fallacy has been pointed out earlier by Cronbach and Snow (1977) when they stated the following:

> A *p* value reached by classical methods is not a summary of the data. Nor does the *p* value attached to a result tell how strong or dependable the particular result is ... Writers and readers are all too likely to read .05 as p(H | E), "the probability that the Hypothesis is true, given the Evidence." As textbooks on statistics reiterate almost in vain, *p* is p(E | H), the probability that this Evidence would arise if the [null] hypothesis is true. Only Bayesian statistics yield statements about p(H | E). (p. 52)

However, not all statistics textbooks, or statistics teachers, make this difference clear. As Shaver (1992) said, though "it seems terribly obvious that a test of statistical significance does not speak directly to causality" (p. 14), statistics textbooks and journal articles continue to perpetuate this fallacy by "concluding that a statistically significant result indicates a treatment effect" (p. 14).

MYTH # 3:
SST Logic Parallels the Logic of Mathematical Proof by Contradiction

Many mathematicians seem to be influenced by the seemingly parallel arguments in a proof by reductio ad absurdum and that of the rejection of the null hypothesis. The essence of the argument involved in a proof by reductio ad absurdum is that a premise that leads logically to an absurd result must be rejected. The most frequently quoted example is the proof that $\sqrt{2}$ is irrational, by assuming that it *is* rational and proceeding to show that such an assumption leads to an absurdity, whereby one has to reject the original assumption of $\sqrt{2}$ being rational.

The parallel argument for SST is that the assumption that H_0 is true leads to an improbable result, and therefore leads logically to the rejection of the assumption that H_0 is true. The improbable result is what is commonly termed a "significant" result. By this argument, a significant result leads to the rejection of H_0.

Falk and Greenbaum (1993) gave various arguments why the analogy between syllogistic reasoning (as exemplified by the reductio ad absurdum proof) and probalistic reasoning is misleading. They emphasised that "disproving the antecedent by denying the consequent" (p. 7), as used in syllogistic reasoning, leads to a confusion between the *improbable* and the *impossible* in probalistic reasoning. They pointed out that a low $P[D \mid H_0]$—that is, a low significance level—does not guarantee a correspondingly low $P[H_0 \mid D]$ to "warrant rejection of H_0" (p. 7). Indeed, even *one* counter-example of a low $P[D \mid H_0]$ but a medium or high $P[H_0 \mid D]$, is sufficient to put the logic of SST in disarray.

They cited a dramatic counter-example from the field of genetic counselling (Pauker & Pauker, 1979), where even if the probability of getting a positive amniocentesis result (indicating Down's syndrome), given the foetus is normal, is only about .005, the probability of a normal foetus, given a positive result is almost .82. In other words, if H_0 = the foetus is normal, H_1 = the foetus is affected, and D = the test result is positive, then $P[D \mid H_0] = .005$, but $P[H_0 \mid D] = .82$, or alternatively, $P[D \mid H_1] = .995$, but $P[H_1 \mid D] = .18$ (see Falk, 1986 and Falk and Greenbaum, 1993, for an elaboration). Falk and Greenbaum (1993) contended that

> although the language of significance tests is not used in reference to medical tests, one employs exactly the same logic when rejecting the hypothesis of normality (in favor of the presence of disease) based on a positive outcome of an 'accurate' test. (p. 8)

It is a sobering thought that the high probability of testing positive for, say, AIDS, given that the person has AIDS, is taken to mean that the probability is also high that the person has that disease, given that a positive test result was obtained.

But, if replicated studies give similar results, then it would be appropriate to say that the conclusion is warranted. Significance levels by themselves, however, are insufficient and erroneous grounds for coming to any conclusions about replicability. So it is the myth of replicability/reliability that I discuss next.

MYTH # 4:
SST Indicates Reliable / Replicable and Valid Results

When a result is reported as statistically significant, many readers and researchers assume that this is an endorsement that such results are "reliable" and therefore have a high probability of being replicated. Some even go so far as to replace the phrase "statistically significant" by the word "reliable" as, for example, when Begg, Armour, and Kerr (1985) reported that "the only reliable effect was the interaction between initial study and statement set, *F*(1,48) = 39.1" (p. 203, cited in Falk and Greenbaum, 1993, p. 12). But, as Carver (1978) said,

> the only valid reason for considering statistical significance is to try to determine whether research results are simply a product of chance and will therefore not be replicable. Yet it is not logical to deduce that if the results are statistically significant, they will replicate, or that if the results are not statistically significant, they will not replicate. (p. 392)

It is easy to see why so many fall prey to this illusion of replicability: when a statistically significant result is interpreted to mean that the result was *not* due to chance, the next logical step is to assume that such results can be reproduced by a non-chance (and usually a human) agency. In other words, the evidence for the research hypothesis (that is, its validity) is supposedly strengthened by the "resultant" low probability of the null hypothesis.

Hence the two fallacies of replicability and validity are mutually dependent. Use of phrases such as "highly significant" for "statistically significant" (see, for example, comments by Carver, 1978, and Slakter et al., 1991) further compounds this illusion. Yet, as Gigerenzer and Murray (1987, p. 25) said, "psychologists seem not to wish to be cured of these illusions." Resistance is not surprising, since giving up such illusions is tantamount to denying that SST can give "objective" and "rigorous" answers to perplexing and uncertain phenomena in education.

The fallacies and misconceptions about SST which I have discussed above are usually made by those who are what could be termed radical proponents of SST, and are certainly not made by all researchers. For instance, Gold (1969), and Winch and Campbell (1969)—who have been labelled conservative and moderate proponents of SST by Carver (1978)—do not believe that SST gives a comprehensive answer to decision-making in research situations. But even the conservative and moderate proponents of SST reveal some serious weaknesses in their arguments (Carver, 1978; Morrison & Henkel, 1969). I will discuss these next.

MYTH # 5:
SST is a Necessary but not a Sufficient Criterion for the Credibility of Research

Proponents of this myth argue that, although statistical significance is not sufficient grounds for deciding whether the research results are important enough, one *has* to establish statistical significance first, before going any further and deciding anything else about the research results. Gold (1969), for example, was emphatic that statistical significance is only a necessary but not a sufficient criterion of substantive (scientific) significance. Carver (1978) commented that "Here, Gold is actually trying to jump two hurdles: evaluate statistical significance first, then evaluate scientific significance" (p. 389). As Morrison and Henkel (1969) said:

> Thus, those who argue for statistical significance as a necessary or minimum criterion for substantive significance for *any* set of cases (Gold, 1969) show an atheoretical orientation in two ways: they disregard the technical requirements of statistical theory, and in addition, provide a criterion of no logical or practical help in building social theory. They do not recognize what is necessary for social theory any better than they recognize what is necessary for statistical theory. (p. 137)

It must be remembered that the statistical inference model necessitates very many replications of studies on randomly selected samples (even non-parametric tests are only meaningful if applied to randomly selected samples). What a theory does, in effect, is to allow one to select and test a limited number of cases under specific conditions, in order to ascertain the corresponding specific expectation of the theory. However, according to Morrison & Henkel (1969),

> when a typical null hypothesis is stated without specifying conditions (as is usual in social sciences), there is no basis for purposive selection of cases to test it, nor much basis for judging the results. In fact, *any* set of conditions constitutes an adequate test for an unconditional hypothesis. (p. 136)

Moreover, the argument that SST is a necessary but not a sufficient condition for substantive significance assumes that substantive significance is directly proportional to the size of the difference or the strength of the relationship between variables. This is atheoretical, because theory aims to predict "the form and strength of empirical relationships *whatever* that strength happens to be." (Morrison & Henkel, 1969, p. 136).

Both Carver (1978) and Morrison and Henkel (1969) gave cogent reasons for doubting the veracity of the "necessary but not sufficient condition" argument propounded by the conservative and moderate users of SST, and the interested reader is strongly recommended to read their papers.

Some Harmful Effects

The overreliance on dichotomous SST decisions often leads to less attention being paid to the design of the study. Indeed, Carver (1978) suggested "a study with results that cannot be meaningfully interpreted without looking at the *p* values is a poorly designed study" (p. 394). Furthermore, Carver stated that SST "is also likely to continue to encourage researchers to investigate hypotheses that are readily tested using research designs that permit neat statistical tests, whether the hypotheses are the most important or not" (p. 397). Thompson (1992), too, has pointed out the inappropriateness of an overreliance on SST in making policy recommendations, as well as to the questionable robustness of ANOVA to "violation of the homogeneity of variance assumptions" (p. 6).

Moreover, "virtually any study can be made to show significant results if one uses enough subjects regardless of how nonsensical the content may be" (Hays, 1974, p. 326). In other words, data can be manipulated to obtain research results which will lead to the rejection of H_0 (which was the original intent of the researcher, anyway), and correspondingly increase the possibility of accepting the research hypothesis.

With regard to the questionable usefulness of H_0 in research, Bakan (1966) pointed out that there are several a priori reasons for disbelieving H_0. Therefore, it would seem to be more fruitful to concentrate on H_1, the research hypothesis, and look for support of this hypothesis, and at the same time determine the direction and size of the mean difference. Because of the emphasis on H_0, the hypothesis of no difference, and on SST in contemporary research, other statistical measures which are not dependent on, say, the size of the sample (e.g., effect size) are either only cursorily studied or used, if at all. Instead, an inordinate amount of time is spent on learning SST by researchers-to-be, "often at the expense of learning about replication designs, confidence intervals, correlation ratios, intraclass correlations, and other effect-size measures" (Carver, 1978, p. 397).

As Coats (1970, p. 6) stated: "Most graduate schools of education still require students to take what may be one of the most *irrelevant learning experiences of their entire educational career.* The requirement is the study of inferential statistics" (italics added). Carver (1978), too, said that "The complete abandonment of statistical significance testing in the training of doctoral students in educational research should be seriously considered" (p. 396). That both Coats' and Carver's view of the educational insignificance of SST is unpopular is evidenced by statements made by even some of those critical of SST. Thompson (1992), for example, talked about "alternatives that may be useful to augment the evaluation of significance testing" (p. 1), and suggested that, rather than abandon SST, "it is useful to have some estimate, albeit a limited one, regarding the probability of a sample result, assuming that the sample came from a population in which the null was true" (p. 7).

Although Thompson's suggested alternatives seem workable, it baffles me how one can argue for SST, something of admittedly limited or trivial value, if this is balanced against the amount of time and money spent on teaching this technique. If only Carver's call for abandoning SST were followed, the seemingly uncritical and mechanical application of SST would be diminished substantially, and educational research based on positivist traditions might get a new lease of life.

Another harmful effect of the continued reliance on SST is the aura of respectability, replicability and generalisability that seems to surround published research which is based on SST. This effectively inhibits actual replications (Bakan, 1966; Sterling, 1959). Though most researchers agree that replications are crucial, "replications are not very common in educational research" (Shaver, 1985a, p. 60). For example, how many doctoral students are encouraged to replicate studies, and how much recognition is given to replicated studies as worthy of "advancing knowledge?"

In addition, the common practice (in the 1960s, especially) of restricting publication to articles showing statistically significant results also prevented non-significant results (which might otherwise indicate non-replicability and might even invalidate previously-published results) from being published (Bakan, 1966). In fact, at one time, Melton (1962), the editor of the *Journal of Educational Psychology,* had openly stated that he considered articles worthy of publication only if they reported a .01 significance level! Also, as Rosenthal (1979) pointed out, not only did editorial practices favour the publication of statistically significant results, but the researchers themselves were unwilling to submit studies showing statistically non-significant results. According to Atkinson, Furlong and Wampold (1982), unwillingness to submit statistically non-significant studies is exacerbated by receiving unfavourable reviews on submission.

Commenting on this, Carver (1978), suggested that "on the contrary, editors should consider rejecting articles that contain this trivial information, just as they presently reject articles that contain raw data" and that "Manuscript referees ... should not allow statistical significance to be interpreted as crucial evidence supporting the stability, reliability, replicability, or importance of the results" (p. 395).

True, hardly any journal editor in the last decade has come out as strongly as Melton (1962) did, in support of reporting at least a .01 significance level, for justifying consideration for publication. However, a cursory glance at most educational journals in the last ten years would reveal the extent to which the practice of reporting significance levels is more the rule than the exception.

Evidently, warnings against relying solely or too heavily on SST, made more than 60 years ago, and repeated frequently over these years seem to have fallen on deaf ears. SST is alive and well, especially as giving it up, for those who have vested interests, is akin to cutting off your nose to spite your face!

Overall, then, the greatest harm done by relying solely on inferences drawn from SST is the mindless ritual to which a research enterprise is reduced. Instead of using informed decisions based on the optimum number of parameters affecting the research (e.g. Shaver, 1985b, 1992; Slakter et al., 1991; Thompson, 1992), SST is being used as the rites of passage into the critical and intellectual world of academia. As Morrison and Henkel (1969) put it, when talking about the arbitrary level of significance commonly used in sociological research to infer from SST, "What we do in sociology surely is much more akin to religion than science and we might as well forget empirical work and get on with the development of more rituals" (p. 137). Similarly, Salsburg (1985), criticised researchers in the medical profession for treating statistics as a religion.

Some Alternatives to SST

Since it would require not just one, but a number of papers to address alternatives to SST in a comprehensive manner, I will just list, and not elaborate on some suggestions. Moreover, others have explained not only what these alternatives are, but also have given examples on how to use them. Thus, it would be remiss of me not to point out where the interested reader can turn for matters concerning these alternatives. A further caveat is in order. I only suggest a few alternatives, and because I agree with Carver (1978) and Falk and Greenbaum (1993) that there is no substitute for replication, I consider the suggested alternatives as only partial solutions to the questions facing researchers.

Below is a brief list of suggested alternatives, with references:

1. Actual replication (e.g., Falk & Greenbaum, 1993; Stevens,1971).
2. Interpreting results based on likelihood of replication: cross-validation (e.g., Thompson, 1989); jacknife (e.g., Daniel, 1989); and bootstrap (e.g., Diaconis & Efron, 1983; Lunneborg, 1987; Thompson, 1988)
3. Confidence intervals (e.g. Rozeboom, 1960; Shaver, 1985a, b).
4. Effect size measures such as: absolute differences, omega squared, eta squared (Hays, 1974), measures of explained variance (Hays, 1981, pp. 289–296, 349–350), proportion of misclassifications (Levy, 1967), efficacy coefficients (Guttman, 1981), simple binomial effect size display (BESD) (Rosenthal & Rubin, 1982), *d* index (Cohen, 1977; Cooper, 1984).
5. Power of the statistical test, $1 - \beta$ = probability of accepting H_1 if H_1 is true (Cohen, 1977).
6. Bayesian posterior probability (Phillips, 1973).

Of these, even those which use the level of significance do so only insofar as to compute other statistical measures and not for purposes of a dichotomous reject-accept decision based solely on the level of significance. And, although most of these alternatives "do not answer the question erroneously believed to be answered by significance tests," they do "provide summary descriptive measures that answer other questions, albeit more related to the appraisal of research results than the one actually answered by the significance test" (Falk, 1986, p. 93). Moreover, most of these measures are free of the "misleading effect the sample size may have on the interpretation of *p* values. Besides, they are commonsensical, easy to comprehend and compute" (Falk, 1986, p. 93). However, there is still no agreement on the usefulness of some of these alternatives. For example, the BESD has been a source of lively debate (e.g., Crow, 1991; McGraw, 1991; Rosenthal, 1990; Strahan, 1991), as has been the use of power analysis (e.g., Cohen, 1990; Chow, 1991; Shaver, 1992). In spite of the suggested alternatives, I feel that actual replication is still the most believable approach to a scientific method (which is what users of SST claim to be employing). However, for non-mathematically inclined researchers, the second alternative above seems to be a promising, if somewhat more daunting, approach (in spite of the availability of relevant computer software).

The ubiquitous use of SST seems to be lessening somewhat nowadays, partly because of the increased acceptance of qualitative studies. Even so, very rarely does one see descriptive statistical measures such as effect size or other alternative statistical tools being utilised in experimental research. The only difference seems to be that the sample size, mean and standard deviation are being reported more often, together with significance levels.

Conclusion

In this paper, I have argued that inferences from SST do not answer the pressing question "Is there a low probability that the research results were due to chance?" I have done this by calling attention to five major misconceptions and fallacies about SST. I have also discussed some undesirable outcomes to educational research which can result from the ritualistic application of inferences drawn from SST, and have suggested that no great loss would accrue from an abandonment of learning SST as a tool for research. Finally, I have suggested some alternatives to SST, and I contend that replications should be the strongest basis of any claim to credibility of research results.

As Stevens (1971) said, "in the long run scientists tend to believe only those results that they can reproduce ... statistical tests of significance, as they are so often miscalled, have never convinced a scientist of anything" (p. 440). If scientists, who usually deal with less confounding variables than their counterparts in mathematics education research, are not convinced by SST, then it would seem inappropriate for the latter to rely on SST. Unfortunately, many mathematics education researchers seem to have been seduced by the mathematical symbolism and apparent logic involved in SST. Although those researchers in education, whose connection with mathematics is not so apparent, may be excused for not examining the taken-for-granted assumptions underlying SST, mathematics education researchers cannot be so excused. Indeed, it is time that mathematics education researchers took a lead in throwing off the shackles of SST, and in showing that credible experimental research results can be reported using statistical tools other than SST, thus allowing for more meaningful replications.

References

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and scientific process: Is there a (statistically) significant relationship? *Journal of Counselling Psychology, 29,* 189–194.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Begg, I., Armour, V., & Kerr, T. (1985). On believing what we remember. *Canadian Journal of Behavioral Science, 17,* 199–214.

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

Chow, S. L. (1991). Some reservations about power analysis. *American Psychologist, 46,* 1088–1089.

Coats, W. (1970). Significant differences: A case against the normal use of inferential statistical models in educational research. *Educational Researcher Newsletter, 21,* 6–7.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1990). Things I've learned so far. *American Psychologist, 45,* 304–312.

Cooper, H. M. (1984). *The integrative research review: A systematic approach.* California: Sage Publications.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Crow, E. L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology?" *American Psychologist, 46,* 1083.

Daniel, L. G. (1989, January). *Use of the jacknife statistic to establish the external validity of discriminant analysis results.* Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, Texas. (ERIC Document Reproduction Service No. ED 305 382).

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist, 42,* 145–151.

Dawes, R. M. (1981). *How to use your head and statistics at the same time, or at least in rapid alternation.* Unpublished manuscript, University of Oregon.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248*(5), 116–130.

Diaconis, P., & Freedman, D. (1981). The persistence of cognitive illusions. *The Behavioral and Brain Sciences, 4,* 333–334.

Factor, L., & Kooser, R. (1981). *Value presuppositions in science textbooks: A critical bibliography.* Galesburg, IL: Knox College.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9,* 83–96.

Falk, R., & Greenbaum, C. W. (1993). *The fallacy of probabilistic modus tollens and the statistical-significance decision.* Paper submitted for publication.

Fisher, R. A. (1960). *The design of experiments,* (7th ed.). Edinburgh: Oliver & Boyd.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gold, D. (1969). Statistical tests and substantive significance. *The American Sociologist, 4,* 42–46.

Guttman, L. (1977). What is not what in statistics. *The Statistician, 26,* 81–107.

Guttman, L. (1981). Efficacy coefficients for differences among averages. In I. Borg (Ed.), *Multidimensional data representations: When and why.* Ann Arbor, MI: Mathesis Press.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1,* 3–10.

Hays, W. L. (1974). *Statistics* (2nd ed.). New York: Holt, Rinehart & Winston.

Hays, W. L. (1981). *Statistics for psychologists* (3rd ed.). New York: Holt, Rinehart & Winston.

Kendall, M. G. (1943). *The advanced theory of statistics.* Vol. 1. New York: Lippincott.

Lesnak, R. J. (1989). Writing to learn: An experiment in remedial algebra. In P. Connolly & T. Vilardi (Eds.),*Writing to learn mathematics and science* (pp.147–156). New York: Teachers College Press.

Levy, P. (1967). Substantive significance of significant differences between two groups. *Psychological Bulletin, 67,* 37–40.

Lunneborg, C. E. (1987). *Bootstrap applications for the behavioral sciences.* Seattle: University of Washington.

McGraw, K. Q. (1991). Problems with the BESD: A comment on Rosenthal's "How are we doing in soft psychology?" *American Psychologist, 46,* 1084–1086.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64,* 553–557.

Morrison, D. E., & Henkel, R. E. (1969). Significance tests reconsidered. *The American Sociologist, 4,* 131–140.

Pauker, S. P., & Pauker, S. G. (1979). The amniocentesis decision: An explicit guide for parents. In C. J. Epstein, C. J. R. Curry, S. Packman, S. Sherman & B. D. Hall (Eds.), *Birth defects: Original article series; Vol. 15. Risk, communication, and decision making in genetic counseling* (pp. 289–324). New York: The National Foundation.

Phillips, L. D. (1973). *Bayesian statistics for social scientists.* London: Nelson.

Rosenthal, R., & Rubin, D. B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775–777.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician, 39*(3), 220–223.

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

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

Shaver, J. P. (1992, April). *What statistical significance testing is, and what it is not.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Slakter, M. J., Yu, Y. B., & Suzuki-Slakter, N. S. (1991). *, **, and ***; Statistical nonsense at the .00000 level. *Nursing Research, 40*(4), 248–249.

Spencer-Brown, G. (1957). *Probability and scientific inference.* London: Longmans.

Stegmuller, W. (1973). *"Jenseits von Popper und Carnap": Die logischen Grundlagen des statitischen Schliessens*. Berlin: Springer.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Stevens, S. S. (1968). Measurement, statistics and the schemapiric view. *Science, 161,* 849–856.

Stevens, S. S. (1971). Issues in psychophysical measurement. *Psychological Review, 78,* 426–450.

Strahan, R. F. (1991). Remarks on the binomial effect size display. *American Psychologist, 46,* 1083–1084.

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. (1988). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. *Educational and Psychological Measurement, 48,* 1129–1135.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counselling and Development, 22,* 2–6.

Thompson, B. (1992). *The use of statistical significance tests in research: Some criticisms and alternatives.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 22, 1992.

Tyler, R. W. (1931). What is statistical significance? *Educational Research Bulletin, 10,* 115–118, 142.

Winch, R. P., & Campbell, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *The American Sociologist, 4,* 140–143.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Author

Rama Menon, School of Science, Nanyang Technological University, National Institute of Education, 469 Bukit Timah Road, Singapore 1025.