Copyright 1998 by the
Mid-South Educational Research Association

RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 43-53

What If There Were No More Bickering About Statistical Significance Tests?

Joel R. Levin
University of Wisconsin – Madison

Questions and concerns are directed to those who advocate replacing statistical hypothesis testing with alternative data-analysis strategies. It is further suggested that: (1) commonly recommended hypothesis-testing alternatives are anything but perfect, especially when allowed to stand alone without an accompanying inferential filtering device; (2) various hypothesis-testing modifications can be implemented to make the hypothesis-testing process and its associated conclusions more credible; and (3) hypothesis testing, when implemented intelligently, adds importantly to the story-telling function of a published empirical research investigation.

From the local pubs to our professional "pubs," everyone in social-science academic circles seems to be talking about it these days. Not that there's anything wrong with talking about it, mind you, even to a more practically oriented crowd such as the readership of this journal. But as with the "gates" of Washington politics on the one coast and the Gates of Washington state on the other, when do we stand up and say "Enough already!"? When do we decide that ample arguments have been uttered and sufficient ink spilled for us to stop talking about it and instead start doing something about it?

The "it," of course, is the "significance test contro-versy" (Morrison & Henkel, 1970), which, in its most extreme form is whether or not conductors/reporters of scholarly research should continue (or even be *allowed* to continue) the time-honored tradition of testing statistical hypotheses. As has been carefully documented in our current forum on the issue in this issue of *RESEARCH IN THE SCHOOLS*, the topic isn't one that just recently arrived on the science scene. Not at all. Eminent statisticians, applied researchers, and just plain folks have been debating the virtues and vices of statistical significance testing for decades, with the debate crescendoing every couple of decades or so – consistent with principles of GC ("generational correctness").

The decade of the 1990s has been a critical one in hypothesis testing's protracted struggle for survival. During this decade especially vitriolic attacks, by

Joel R. Levin is a professor of educational psychology at the University of Wisconsin. Correspondence concerning this article should be addressed to Joel R. Levin, Department of Educational Psychology, 1025 W. Johnson St., University of Wisconsin, Madison, WI 53706 (E-mail address: LEVIN@MACC.WISC.EDU).

especially viable attackers, in especially visible outlets (e.g., Cohen, 1990, 1994; Kirk, 1996; Schmidt, 1996), have been mounted for the greater good of God, country, and no more significance testing! Even more critically for the life-and-death struggle, in the 1990's we also witnessed the first formal establishment of task forces and committees representing professional organizations [e.g., the American Psychological Association (APA), the American Educational Research Association (AERA), the American Psychological Society (APS)] to study the "problem" and make recommendations. As the deliber-ations of such task forces have proceeded apace, so have the spoken and written words: for example, in semi-civilized debates at professional meetings [e.g., "Significance tests: Should they be banned from APA journals?" (APA, 1996); "Should significance tests be banned?" (APS, 1996); "A no-holds-barred, tag-team debate over the statistical significance testing contro-versy" (AERA, 1998)] and in the most comprehensive, most indispensable, source on the topic, the edited volume *What if there were no significance tests*? (Harlow, Mulaik, & Steiger, 1997; reviewed by Levin, 1998, and Thompson, 1998).[1]

In the typical argument scenario, hypothesis testing is cast as the "bad guy," the impeder of all scientific progress. The prosecution prosecutes the accused, and then the defense defends. That is the basic approach taken in Harlow et al.'s (1997) four focal chapters ("The Debate: Against and For Significance Testing"), as well as in the recent professional meeting set-to's. As each piece of hypothesis-testing evidence is trotted out for public display, the typical juror-consumer goes through a "good point, that sounds reasonable, I hadn't thought of that" self-dialogue before deciding whether to convict or acquit, or just to quit and retreat to his/her original position on the subject.

Comments and Questions Related to the Present Articles

A similar structure and sequence of events are witnessed in the present collection of three essays. The "bad guy, good guy" script is closely followed, with each essay providing informative backgrounding, coherent evidence, and a convincing closing argument in the form of practical suggestions and proposed solutions. At the same time, even though the Editors of *Research in the Schools* have striven to be impartial and maintain a balance of perspectives here, the fact that two of the essays are clearly hypothesis-testing indictments whereas only one supports the process indicates that the present scales of justice were tipped a priori toward conviction. Given this unfair state of affairs and not knowing in advance the substance of the other critics' critiques, I can be "up front" in my admission of evening out the imbalance with the comments I am about to make.[2]

All the authors of the present articles cite relevant literature in a scholarly fashion and then proceed to make their case. As a reminder of what those cases are: (a) Nix and Barnette (1998) nix hypothesis testing in favor of a number of more thought-to-be informative alternatives to it (the provision of effect sizes, confidence intervals, replication, meta-analyses); (b) Daniel (1998) basically concurs and then goes on to recommend specific journal editorial-policy measures that could be implemented to effect those changes; and (c) McLean and Ernest (1998) disagree with the fundamental assertion about hypothesis testing's inutility, arguing essentially that it has an important "time and place" (Mulaik, Raju, & Harshman, 1997) in the scientist's analytic arsenal.

Although I have found it unwise to argue with people on matters of politics, religion, and their convictions about hypothesis testing, I will nonetheless attempt to do so by commenting on selected specifics in the three focal articles, in no particular order. Included in my comments are a number of questions that the articles evoked, the responses to which I look forward to reading in the authors' rejoinders. With the exception of Nix and Barnette's discussion of "research registries" (which I found to be a useful notion that should be given serious consideration by social scientists), the case against hypothesis testing introduces all the usual suspects. In that the present authors have examined these suspects in a generally commendable fashion, I will do my best to cross-examine them. In addition to being invited to serve as a commentator on these articles, I was encouraged to get in my own "two bits worth." And so I shall, beginning with a confession: Because of my previously professed "pro" position in the hypothesis-testing debates, I apolo-gize in advance for disproportionately carping and sniping more at the "con" positions of Nix-Barnette and Daniel.

Hypothesis-Testing Fever/Furor

Considerable issue can be taken with something that Nix and Barnette claim early on, namely, that "the informed stakeholders in the social sciences seem to be abandoning NHST . . ." As one who considers himself to be an informed stockholder, I'd be curious to learn to whom Nix and Barnette are referring, on what survey or other supporting reference their claim is made, and exactly how prevalent this abandonment is. One has to wonder: If the perniciousness of hypothesis testing is so pervasive, then why has APA's elite task force recommended that the practice *not* be abandoned, but rather supplemented and improved by many of the same enhancements that are mentioned in the present exchange (*viz*., effect magnitude measures, confidence intervals, replications, and meta-analysis)?

It is understandable that much, if not most, of what Daniel decries and prescribes has been decried and prescribed before. It is understandable because: (a) Daniel draws heavily from the words and work of Bruce Thompson (11 references and counting); and (b) Daniel, as a Thompson collaborator (Thompson & Daniel, 1996a, 1996b), is undoubtedly quite familiar with that corpus. Prominent in Daniel's list of hypothesis-testing do's and don'ts are Thompson's (e.g., 1996) "big three" recommended editorial policy "requirements" for authors of empirical studies – namely, that authors must always: (a) modify the word "significant" with "statistically," in reference to hypothesis tests; (b) include explicit effect-size information; and (c) provide some form of outcome "replicability" evidence.

"Significance" Testiness

Such proposed editorial policy changes are sensible enough and I clearly support the spirit – though not the letter – of them (e.g., Levin & Robinson, in press; Robinson & Levin, 1997). What is difficult to support are requirements that take away certain freedoms of author style and expression; in particular, when editorial policy is only half a vowel away from turning into editorial police. For example, when addressing a professional audience with a shared understanding of technical terminology, why should an author be forced into using stilted, reader-unfriendly, language (e.g., "The two correlations are each statistically significant but not statistically significantly different from one another.")? In a Results section where statistical hypotheses are being tested, there can be no misunderstanding what the word "significant" does or does not mean; the context disambiguates the concept. On the other hand, if an author who detects an effect that is significant statistically (e.g., a significance probability of p = .01) but insignificant practically (e.g., a standardized difference in means represented by a Cohen's d of .01) goes on to talk about the effect with reckless hyperbole, then, yes, that author should be shot at sunrise – or at least appropriately chastised.[3]

Effect-Size Defects?

Speaking of talking, the just-mentioned confusion represents a profound mismatch between an author's evi-dence and his/her words, stemming from a preoccupation with statistical significance at the expense of taking into account the magnitude of the obtained effect (which in the d = .01 case was minuscule). However, I have problems with the other side of the "nouveau" editorial-policy- recommendations coin regarding effect-size reporting as well. I will mention a few such problems, none of which is noted either by Daniel or by Nix and Barnette.

First, and even though I am all for including effect sizes as ancillary evidence of outcome importance, it has been pointed out previously (Levin & Robinson, in press; Robinson & Levin, 1997) that there are extremists in the mandatory effect-size camp (including journal reviewers and editors) who advocate reporting and concentrating on effect sizes *only* (i.e., without accompanying statistical/probabilistic support). This practice is absurdly pseudoscientific and opens the door to encouraging researchers to make something of an outcome that may be nothing more than a "fluke," a chance occurrence. Without an operationally replicable screening device such as statistical hypothesis testing, there is no way of separating the wheat (statistically "real" relationships or effects) from the chaff (statistically "chance" ones), where "real" and "chance" are anchored in reference to either conventional or researcher-established risks or "confidence levels." McLean and Ernest's description of Suen's (1992) "overbearing guest" analogy is especially apt in this context.[4]

Examples of the seductive power of large observed effect sizes that are more than likely the result of chance outcomes are provided by Levin (1993) and Robinson and Levin (1997). In its extreme form, effect-size-only reporting degenerates to strong conclusions about differential treatment efficacy that are based on comparing a single score of one participant in one treatment condition with that of another participant in a different condition. Or, even more conveniently and economically (i.e., in situations where time and money are limited), how about conclusions from a "what if" meta-experiment in which scores of two *imaginary* participants are compared (N = 0 studies)? The latter tongue-in-cheek situation aside, consider the following proposition:

> Suppose that Aladdin's genie (Robin Williams?!) pops out of the lamp to grant you only *one* forced-choice wish in relation to summarized reports of empirical research that you will read for the rest of

> your lifetime: You can have access to either a statistical-significance indicator of the reported findings or a practical-significance index of them, but not both (and no sample-size information can be divulged). Which would you choose?

Personally speaking, it would be painful to have to choose only one of these mutually exclusive alternatives. Based on the aforementioned "chance" and "seductive effect size" arguments, however, I think that a strong case can be made for statistical over practical significance. McLean and Ernest's chance-importance-replicability trichotomy represents a nice way of thinking about the problem, with an assessment of the findings' nonchanceness and replicability each given priority over importance. At the same time, l heartily endorse Nix and Barnette's statement, "We would like to see a situation where all studies that were adequately designed, controlled and measured would be reported, regardless of statistical significance." In fact, I am quite sympathetic with others who have called for manuscript reviews and editorial decisions based on just a study's rationale, literature review, and methods and procedures, in the form of a research proposal – with the associated outcomes and data analyses not included until an editorial decision has been reached (e.g., Kupfersmid, 1988; Levin, 1997; Walster & Cleary, 1970a).

So you want to change the world? Nix and Barnette, as well as Daniel, make it sound as though the research world will be a far better place when the hypothesis- testing devil is ousted by the effect-size angel. In my opinion, that is not a fair assumption, as effect-size calculating and reporting are subject to the same "bias" criticisms inherent in familiar "how to lie with statistics" treatises. How to lie with effect sizes? Levin and Robinson (in press) have noted how researchers can select from any number of conventional effect-size measures (including both more and less conservative variants of the indices listed in Nix and Barnette's Table 1, among others) to make the preferred case for their own data. Another problem associated with relying on commonly calculated effect sizes alone is illustrated in the following hypothetical example.

Suppose that an investigator wants to help older adults remember an ordered set of ten important daily tasks that must be performed (insert and turn on a hearing aid, take certain pills, make a telephone call to a caregiver, etc.). In a sample of six elderly adults, three are randomly assigned to each of two experimental conditions. In one condition (A), no special task instruction is given; and in the other (B_1), participants are instructed in the use of self-monitoring strategies. Following training, the participants are observed with respect to their success in performing the ten tasks. As can be seen in the first two columns of Table 1, the average number of tasks the participants correctly remembered to perform was 1.33 and 3.33 for the no-instruction (A) and self-monitoring (B_1) conditions, respectively. For the data provided in Table 1, it can be determined that the "conditions" factor accounts for a hefty 82% of the total variation in task performance (i.e., the squared point-biserial correlation is .82, which for the two-sample case, is equivalent to the sample ç^2). Alternatively, the self-monitoring mean is 3-½ within-group standard deviations higher than the no-instruction mean (i.e., Cohen's *d* is 3.5). From either effect-size perspective (ç^2 or *d*), certainly this represents an impressive treatment effect, doesn't it? Or does it?

Table 1
Hypothetical Data Illustrating Equivalent Standardized Effect Sizes (Condition B Versus Condition A) With Vastly Different Practical Implications

	Condition A	Condition B_1	Condition B_2
	1	3	5
	1	3	8
	2	4	10
M	1.333	3.333	7.667
SD	.577	.577	2.517

Suppose that instead of self-monitoring training, participants were taught how to employ "mnemonic" (systematic memory-enhancing) techniques (B_2) – see, for example, Carney & Levin (1998) – with the results as indicated in the third column of Table 1. The corre-sponding B_2 mean is 7.67 correctly remembered tasks and a comparison with no-instruction Condition A surprisingly reveals that once again, the conditions factor accounts for 82% of the total variation in task performance (equivalently, *d* again equals 3.5).[5] Thus, when expressed in standardized/relative terms (either ç^2 or *d*), the effect sizes associated with the two instructional conditions (B_1 and B_2) are exactly the same, and substantial in magnitude. Yet, when expressed in absolute terms and with respect to the task's maximum, there are important differences in the "effects" of B_1 and B_2: Increasing participants' average performance from 1.33 to 3.33 tasks remembered seems much less impressive than does increasing it from 1.33 to 7.67. Helping these adults remember an average of only 3 of their 10 critical tasks might be regarded as a dismal failure, whereas helping them remember an average of almost 8 out of 10 tasks would be a stunning accomplishment. Yet, the conventional effect-size measures are the same in each case.[6]

How, then, not to lie with effect sizes? To borrow from Cuba Gooding, Jr.'s character in the film, *Jerry Maguire*: Show me the data! Show me, the reader, "sufficient" data (American

Psychological Association, 1994, p. 16) either in raw (preferably) or in summary form. Then, let me, the reader, decide for myself whether a researcher's particular finding is educationally "significant" or "important," with respect to the standards that I regard as "significant" or "important" (see also Prentice & Miller, 1992).

Lack-of-confidence intervals. Briefly noted here are other suggested alternatives to hypothesis testing that are briefly noted by Daniel, as well as by Nix and Barnette. These include the inclusion of confidence intervals and meta-analyses, both of which are signature recom-mendations of Schmidt and Hunter (e.g., 1997). As far as the former are concerned, it is well known that one can simply slap a standard error and degree of confidence on an effect size and build a confidence interval *that is equivalent to testing a statistical hypothesis* (but see McGrath, 1998). Schmidt, Hunter, and their disciples, however, eschew that particular application and instead encourage researchers to select two or three or four or five degrees of confidence (e.g., 99%, 95%, 90%, 80%, 70%) and then build/display two or three or four or five corresponding confidence intervals. Well and good, but how is the researcher or reader to interpret these varying-degrees-of-confidence intervals, and what is one to conclude on the basis of them (e.g., when a 95% interval includes a zero treatment difference but a 90% interval does not)? How much confidence can one have in such subjective nonsense?

I never met a meta-analysis . . . Concerning meta-analyses: I have nothing against them. They can be extremely valuable literature-synthesis supplements, in fact. Yet, their purpose is surely quite different than that of an individual investigator reporting the results of an individual empirical study, especially when the number of related studies that have been previously conducted are few or none. Alas, what's a poor (graduate-student or otherwise) single-experiment researcher to do (Thompson, 1996)? Of course, if the logical corollary to the meta-analysis argument is that no single-experiment reports should be published in empirical journals as we currently know them, then count me in! I strongly endorse the recommendation that replications and multiple-experiment "packages" comprise an essential aspect of a researcher's LPU ("least publishable unit") – see, for example, Levin (1991, p. 6).

Robust Conclusions Versus Replicated Outcomes

There's something about "replication" in two of the present articles with which I take issue. That something is a restatement of Thompson's (1993) view that data-analysis strategies such as cross-validation, boot-strapping, and jackknifing "indicate the likelihood of replication" (Nix and Barnette) or "may provide an estimate of replicability" (Daniel). For readers not in the know and who might be misled by such semantic twists, allow me to elaborate briefly. A "replication" defined by corroborating analyses based on alternative slices or samples of the same data – which applications of the just-mentioned statistical procedures attempt to do (see, for example, Efron & Gong, 1983) – is nice for establishing *the robustness of a single study's conclusions* (Thompson's "internal" replication). However, that type of "replication" is neither as impressive nor as imperative for the accumulation of scientific knowledge as is a "replication" defined by *an independently conducted study* (i.e., a study conducted at different sites or times, with different specific participants and operations) *that yields outcomes highly similar to those of the original study* (Thompson's "external" replication) – see, for example, Neuliep (1993) and Stanovich (1998). Even to suggest that researchers should be satisfied with the former, by rationalizing about researchers' diminished physical or fiscal resources (as both Thompson and Nix and Barnette do), is not in the best interest of anyone or anything, and especially not in the best interest of educational research's credibility within the larger scientific community.

What if there were no more bickering about significance tests? Conclusion robustness itself is a matter of no small concern for researchers, for outcome "credibility" (Levin, 1994) and generalizability depend on it. Yet, because of the excessive "heat" (Thompson, in press) being generated by hypothesis-testing bickerers, little time is left for shedding "light" on how to enhance the conclusion robustness of educational and psychological research. In addition to the methodological adequacy of an empirical study (e.g., Levin, 1985; Levin & Levin, 1993; Stanovich, 1998), the credibility of its findings is a function of the study's "statistical conclusion validity" (Cook & Campbell, 1979), which in turn encompasses a consideration of the congruence between the statistical tools applied and their associated distributional assump-tions. Reviews of the literature indicate that precious little attention is being paid by researchers and journal referees alike to that congruence: Statistical tests are being mindlessly applied or approved even in situations where fundamental assumptions underlying them are likely grossly violated (e.g., Keselman et al., in press; Wilcox, 1997).[7] Bickering time spent on significance testing is also time away from considering other critical conclusion-robustness issues, including particularly those associated with the pervasive educational research realities of: nonindependent sampling, treatment, and testing units; random (as opposed to fixed) treatment factors; longitudinal and other multivariate designs, among others (e.g., Clark, 1973; Kratochwill & Levin, 1992; Levin, 1992a; Raudenbush & Bryk, 1988; Willett & Sayer, 1994). Accompanied or not by significance testing per se, such statistical issues remain properly "significant."

That concludes my comments on the "big issues" addressed by the three focal articles in this issue of *Research in the Schools*. Before concluding with a few additional big issues of my own, I will address several misleading and erroneous statements that appear in the present articles. Though not of the magnitude of the issues just discussed, these statements are nonetheless sufficiently distressing that they should not go unmentioned.

Misleading and Erroneous Assertions in the Present Articles

It is bad enough when *consumers* of research reports are uninformed with respect to the methods and meanings of the data analyses reported (as has been claimed, for example, with respect to the hypothesis-testing term "significant"). Even worse is when *researchers/authors* are misinformed with respect to those methods or meanings. But worst of all is when *critics* of data-analytic practices dangerously mislead or make erroneous assertions regarding those practices – and particularly when the words "misuse and misinterpretation" are featured in the title of a critic's critique (as in Daniel's article, for example).

Sample size and statistical power. To wit, consider Daniel's comments about the components of an *F*-test of mean differences, which I quote [with numbers added for convenience in referencing]:

> . . . the mean square for the error term will get smaller as the sample size is increased [1] and will, in turn, serve as a smaller divisor for the mean square for the effect [2], yielding a larger value for the *F* statistic [3]. In the present example (a two-group, one-way ANOVA), a sample of 302 would be five times as likely to yield a statistically significant result as a sample of 62 simply due to a larger number of error degrees of freedom (300 versus 60) [4].

What a misrepresentation of the *F*-test and its operating characteristics! The error mean square (*MSE*) is an unbiased estimator of the population variance (ó^{2}) that is not systematically affected by sample size. What increasing sample size does is to reduce the sampling variability associated with each condition's mean, which results in increased variability *among* those means, which in turn increases the mean square between conditions (*MSB*) in the *F*-test's *numerator*. Propositions [1] and [2] are therefore false, which invalidates proposition [3]. Proposition [4] is not true as a result of the preceding illogic.

It is also false as a consequence of Daniel's stated "larger number of error degrees of freedom." Again, larger sample sizes increase statistical power by decreasing the sampling variability associated with each condition's mean, which operates to increase the variability among those means. None of this works auto-matically to increase the *F*-statistic by a constant amount, however, as is asserted by Daniel (e.g., "by five times"), *unless* it is also stated that *all else (except sample size) is held constant* – which includes the value of *MSE* and the means for each condition (all of which are statistics that will vary unsystematically with changes in sample size). To give the impression that merely increasing sample size *guarantees* a larger *F*-ratio, as Daniel and others imply, is unfortunate because it simply is not true.

Show you the data? Don't press the issue. I could come up with dozens – if not hundreds, thousands, or zillions, if I had the time and temperament – of examples from actual empirical studies, many from my own substantive research program, where an *F*-ratio based on small sample sizes (calculated, for example, early in the data-collection process) becomes *smaller* when based on larger or final sample sizes.

Some of Nix and Barnette's assertions about statistical power and a study's publishability are similarly misleading. First, the authors state that the problem is of special concern in educational research, where ". . . effect sizes may be subtle, but at the same time, may indicate meritorious improvements in instruction and other class-room methods." If instructional improvements are indeed "meritorious," then: (a) effect sizes will not be "subtle;" and (b) even with modest sample sizes, statistical signifi-cance will follow. Second, many readers are likely to be misled by the authors' statements that "reliability . . . can be controlled by reducing . . . sampling error" and "the most common way of increasing reliability . . . is to increase sample size." Reducing *sampling* error or increasing *sample* size (the number of participants) does not increase reliability. Reducing *measurement* error or increasing *test* size (the number of items) does. Increasing sample size increases the power or sensitivity of a statistical test, however.

Errors and effect sizes. Nix and Barnette also state that in a hypothesis-testing context, "errors can be due to treatment differences." This statement will come as news to many and deserves some elaboration. In the section entitled "Misunderstanding of *p* values," the authors caution that "differences of even trivial size can be judged to be statistically significant when sampling error is small (due to a large sample size and/or a large effect size) . . ." How can a difference be simultaneously "trivial" and "large?" Read that sentence again. Later in the same section, the authors argue that researchers should "con-tinue to determine if the statistically significant result is due to sampling error or due to effect size." The imprecisely worded statement may lead an uninitiated reader to believe that it is actually possible for a researcher to make such a precise either-or determination, when it is not. In Nix and

Barnette's section, "Interpreting effect size," the impression is given that the various U measures are separate/unrelated, when in fact they are alternative ways of thinking about the same outcome – just as is converting d (a standardized difference in means) to r (the correlation between treatment and outcome), something that was left unsaid. Omitted from a subsequent paragraph is the caution that comparing single-study effect sizes with composite effect sizes can be grossly misleading unless all treatments in question are evaluated relative to functionally equivalent "control" groups (see also Levin, 1994).

Hypothesis Testing as a Meaningful, Memorable Process

In this section I will provide a few personal thoughts about statistical hypothesis testing and its rightful role in the analysis and reporting of empirical research in education and psychology.

Dump the Bathwater, Not the Baby...

No, statistical hypothesis testing, as is generally practiced, is not without sin. I too oppose mindless (e.g., Cohen's, 1994, "rare disease" scenario; Thompson's, 1997, "reliability/validity coefficient testing" criticism) and multiple (e.g., testing the statistical significance of all correlations in a 20 x 20 matrix) manifestations of it. Such manifestations surely portray the practice of hypoth-esis testing at its worst. More forethought and restraint on the part of researchers would likely help to deflect much of the criticism concerning its misapplication.

Absent in each of the present articles' proposed replacement therapies for traditional statistical hypothesis testing are *alternative hypothesis-testing therapies themselves* – which I have referred to generically as "intelligent" hypothesis-testing practices (Levin, 1995) and which have been articulated in a set of ideal principles (Levin, 1998). The overarching premise is that statistical hypothesis testing can be a valuable decision-making tool, if implemented in conjunction with a researcher's a priori (i.e., prior to data collection) planning, specification, or determination of:

- a select number of carefully developed (prefer-ably, theory-based) hypotheses or predictions
- a statistical test or tests that validly and parsimoniously assess those hypotheses
- Type I error probabilities that are adequately controlled
- magnitudes of effects that are regarded as substantively "important," along with their associated probabilities of detection
- magnitudes of effects that are regarded as substantively "trivial," along with their associated probabilities of nondetection
- sample sizes that directly follow from these specifications.

The more of these ingredients that are incorporated into the hypothesis-testing process, the more intelligent and informative is that process.

Effects that emerge as statistically significant as a result of intelligent hypothesis testing should be supple-mented by ancillary "practical significance" information, including effect sizes (based on relative and/or absolute metrics), confidence intervals, and even – heaven forbid! – more "qualitative" assessments of treatment efficacy (e.g., experimenter observations and participant self-reports). *The* most important supplement to this statistical basis for scientific hypothesis confirmation is evidence accumulation, initially through empirical replications (Levin's, 1995, "A replication is worth a thousandth p-value.") and ultimately through literature syntheses (which include the tools of meta-analysis).

In contrast to the anti-hypothesis-testing reforms in the graduate-level statistics courses taught at Michigan State (alluded to by Nix and Barnette), UW-Madison colleague Ron Serlin and I attempt to impart intelligent hypothesis-testing practices to our students. In addition to simply teaching and writing about the potential of such improvements to statistical hypothesis testing (e.g., Levin, 1985, 1997; Seaman & Serlin, in press; Serlin & Lapsley, 1993), we also attempt to practice these preachings in our substantive research investigations. For example, Ghatala and Levin (1976, Exp. 2) adapted Walster and Cleary's (1970b) procedure for determining "optimal" sample sizes to distinguish between substantively important and trivial effects based on acceptable Type I error control and statistical power. Similarly, I convinced a former student to incorporate components of "predicted pattern testing" (Levin & Neumann, in press) to provide stronger, more sensible, tests of his theoretically based predictions – see Neumann and DeSchepper (1991, Exp. 3).

To present a case for a place for intelligent statistical hypothesis testing in educational research, I invite you to imagine the following seemingly far-from-educational- research situation:

> Suppose that you are a medical doctor, whose life work is to keep people alive. A particular patient fits a profile for being "at risk" for developing some dangerous abnormality. You need to make a decision, based on a simple screening test, whether or not to proceed to more extensive/expensive testing. For patients with this kind of "at risk" profile, the screening test is known to have a 90% chance of identifying those who have the abnormality to some substantial degree, a 5% chance of identifying those who have the abnormality only to some very minimal

> degree, and a 1% chance of identifying those who do not have the abnormality at all.[8]

Based on the preceding information, does it seem reasonable to you, as a responsible doctor, to use the screening test as a basis for making a decision about whether or not to proceed to the next phase of evaluation? It does to me.

> OK, now suppose that you are an educational researcher whose life work is to study ways of improving the academic performance of "at risk" students. You have developed a literature- guided intervention for "at risk" middle-school students and you want to assess its effectiveness by comparing the end-of-year educational achievement of students who receive the intervention and those who do not (randomly determined). If the intervention produces a substantial difference in average achievement between the two groups (operationalized as d = 1.00), you want to have a 90% chance of detect-ing it; if it produces a minimal difference (d = .25), you only want a 5% chance of detecting it; and if there is no difference at all (d = .00), you are willing to tolerate a risk of 1% of falsely detecting that. Adapting the Walster-Cleary (1990b) approach, for example, indicates that the just-specified parameters and probabilities are satisfied if 32 students are randomly assigned to each of the two conditions (intervention and no intervention).

Based on the preceding information, does it seem reasonable to you, as a responsible educational researcher, to perform a statistical test as a basis for making a decision about the intervention's potential? It does to me – and especially because the situation just described incorporates the earlier listed intelligent hypothesis-testing ingredients. I certainly do not claim this hypothetical educational hypothesis-testing example to represent a detail-by-detail correspondence with the equally hypothetical medical screening-test example. Rather, it constitutes a close enough analogy that takes us through a similarly sensible decision-making process.

. . . And Now the Rest of the Story

I conclude my remarks with a story relevant to our discussion of hypothesis testing's proper place on the empirical research plate.

> It is a dark and stormy night. A shot rings out in the presidential palace. A body slumps and falls to the ground, dead. A one-armed man is seen fleeing the scene. Inspectors Poirot and Clouseau are called in to investigate. Poirot determines that the deceased is the president's lover. Clouseau notices a charred sheet of paper in the fireplace. He picks it up. "Ooooohh, it's still hot!" he yelps, but is nonetheless able to discern some scribblings on the paper. "Zoot, alors, I have it! And I know precisely how it happened!" Clouseau crows. He continues: "The murderer is . . . [pause] . . . the president's men . . . [pause] . . . or possibly it's the one-armed man . . . [pause] . . . or perhaps it's even the president herself . . . [pause] . . . I haven't a clew!"

Hey, c'mon, who dunnit? Tell us the rest of the story. Inquiring minds want to know!

So you want to know the ending? Let me tell you a different story. Somewhere along the academic trail I had an epiphany about reports of empirical research in scholarly journals (at least those in the fields of psychology and education): In addition to describing what was done, how it was done, and what was found, a journal article should "tell a story." I'm not using "story" in the fictional sense here, but rather as true to life and justifiable on the basis of the study's specific operations and outcomes. Telling a story, with a clever "hook" and memorable take-home message, represents a key land-mark on the publication highway (e.g., Kiewra, 1994; Levin, 1992b; Sternberg, 1996). It is something that editors usually demand, reviewers seek, and readers require. A study without a meaningful, memorable story is generally a study not worth reporting. In certain situations, and in light of my earlier comments, incorporating one or more additional experiments into a one-experiment study often helps to breathe life into an otherwise moribund article.

Exactly what does any of this have to do with our current hypothesis-testing discussion? I believe that an invaluable, though heretofore overlooked, function of statistical hypothesis testing (especially if implemented intelligently) is to assist an author in developing an empirical study's story line and take-home message. Just as with the preceding Clouseauian fantasy with its inconclusive conclusion (or its invent-your-own ending), an empirical research article without an evidence-based conclusion is not likely to satisfy either the reader's affective (interest, enjoyment) or cognitive (under-standing, memory) processes. We human animals seek to extract some form of order from the chaos in the world around us; we are all "meaning makers." As consumers of scientific research, we seek to do the same from the jumble of theory, methods, and results that are provided in a journal article. In my opinion, selective, planful statistical hypothesis testing can help one extract order from chaos, not just in the "chance-finding filtering" sense, but in the sense of cementing as firm a conclusion as can be made from the evidence presented until a critical replication-attempting study comes along. I additionally believe that hypothesis testing is much better suited to that cementing task than are other proposed individual alternatives for summarizing the results of single-study investigations, including the provision of effect sizes (are they real?) and multiple-confidence-level confidence intervals (which one do you prefer?).[9]

I could go on about the story-telling function of journal articles and hypothesis testing, but I think you get the idea. As for stories, what's the take-home message of *this* article? There are actually three take-home messages, each enumerated in the Abstract. If you're interested, go back and (re)read them. That, of course, is what journal abstracts are supposed to summarily convey: the "bottom line" of one's work.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.

Carney, R. N., & Levin, J. R. (1998). Mnemonic strategies for adult learners. In M. C. Smith & T. Pourchot (Eds.), *Adult learning and development: Perspectives from educational psychology* (pp. 159-175). Mahwah, NJ: Erlbaum.

Clark, H. H. (1973). The language-as-fixed-effects fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335-359.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*, 997-1003.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand McNally.

Daniel, L.G. (1998). Statistical significance testing: A Historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *RESEARCH IN THE SCHOOLS*, *5*(2), 23-32.

Derry, S., Levin, J. R., & Schauble, L. (1995). Stimulating statistical thinking through situated simulations. *Teaching of Psychology*, *22*, 51-57.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, *37*, 36-48.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, *23*, 89-105.

Frick, R. W. (1995). *Using statistics: Prescription versus practice*. Unpublished manuscript, Department of Psychology, State University of New York at Stony Brook.

Ghatala, E. S., & Levin, J. R. (1976). Phenomenal background frequency and the concreteness/imagery effect in verbal discrimination learning. *Memory & Cognition*, *4*, 302-306.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), *Review of Research in Education* (Vol. 5, pp. 351-379). Itasca, IL: Peacock.

Harlow, L. L., Mulaik, S. A., & Steiger, J. A. (Eds.). (1997). *What if there were no significance tests*? Mahwah, NJ: Erlbaum.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (in press). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*.

Kiewra, K. A. (1994). A slice of advice. *Educational Researcher*, *23*(3), 31-33.

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

Kratochwill, T. R., & Levin, J. R. (Eds.). (1992). *Single-case research design and analysis: New developments for psychology and education*. Hillsdale, NJ: Erlbaum.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist*, *43*, 635-642.

Levin, J. R. (1985). Some methodological and statistical "bugs" in research on children's learning. In M. Pressley & C. J. Brainerd (Eds.), *Cognitive learning and memory in children* (pp. 205-233). New York: Springer-Verlag.

Levin, J. R. (1991). Editorial. *Journal of Educational Psychology*, *83*, 5-7.

Levin, J. R. (1992a). On research in classrooms. *Mid-Western Educational Researcher*, *5*, 2-6, 16.

Levin, J. R. (1992b). Tips for publishing and professional writing. *Mid-Western Educational Researcher*, *5*, 12-14.

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

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review*, *6*, 231-243.

Levin, J. R. (1995, April). *The consultant's manual of researchers' common stat-illogical disorders*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Levin, J. R. (1997). Overcoming feelings of power-lessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging*, *12*, 84-106.

Levin, J. R. (1998). To test or not to test H_0? *Educational and Psychological Measurement*, *58*, 313-333.

Levin, J. R., & Levin, M. E. (1993). Methodological problems in research on academic retention programs for at-risk minority college students. *Journal of College Student Development*, *34*, 118-124.

Levin, J. R., & Neumann, E. (in press). Testing for predicted patterns: When interest in the whole is greater than in some of its parts. *Psychological Methods.*

Levin, J. R., & Robinson, D. H. (in press). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist*, *53*, 796-797.

McLean, J. E., & Ernest, J. M. (1998). The Role of statistical significance testing in educational research. *Research in the Schools*, *5*(2), 15-22.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In Harlow, L. L., Mulaik, S. A., & Steiger, J. A. (Eds.). *What if there were no significance tests*? (pp. 65-115). Mahwah, NJ: Erlbaum.

Neuliep, J. W. (Ed.). (1993). Replication research in the social sciences. Special issue of the *Journal of Social Behavior and Personality*, *8*(6).

Neumann, E., & DeSchepper, B. G. (1991). Costs and benefits of target activation and distractor inhibition in selective attention. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *17*, 1136-1145.

Nix, T.W., & Barnette, J.J. (1998). The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. *RESEARCH IN THE SCHOOLS*, *5*(2), 3-14.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, *92*, 766-777.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160-164.

Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, p. 423-475). Washington, DC: American Educational Research Association.

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

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of signifi-cance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. A. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Seaman, M. A., & Serlin, R. C. (in press). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods.*

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues* (pp. 199-228). Hillsdale, NJ: Erlbaum.

Stanovich, K. E. (1998). *How to think straight about psychology* (5th ed.). New York: Longman.

Sternberg, R. J. (1996). *The psychologist's companion: A guide to scientific writing for students and researchers* (3rd ed.). Cambridge, UK: Cambridge University Press.

Suen, H. K. (1992). Significance testing: Necessary, but insufficient. *Topics in Early Childhood Special Education*, *12*, 66-81.

Thompson, B. (1993). The use of statistical signifi-cance tests in research: Bootstrap and other alter-natives. *The Journal of Experimental Education*, 61(4), 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26-30.

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

Thompson, B. (1998). Review of *What if there were no significance tests*? *Educational and Psychological Measurement*, *56*, 334-346.

Thompson, B. (in press). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*.

Thompson, B., & Daniel, L. G. (1996a). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, *56*, 197-208.

Thompson, B., & Daniel, L. G. (1996b). Seminal readings on reliability and validity: A "hit parade" bibliography. *Educational and Psychological Measurement*, *56*, 741-745.

Walster, G. W., & Cleary, T. A. (1970a). A proposal for a new editorial policy in the social sciences. *The American Statistician*, *24*, 16-19.

Walster, G. W., & Cleary, T. A. (1970b). Statistical significance as a decision-making rule. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology* (pp. 246-254). San Francisco: Jossey-Bass.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*, 363-381.

Footnotes

[1] The authors of the present exchange can certainly be excused for their limited reference to the Harlow et al. volume, as it likely was released only after earlier versions of the current articles had been written and submitted.

[2] Psst! It should be a secret to nobody that I am a staunch hypothesis-testing defender (e.g., Levin, 1993, 1998; Robinson & Levin, 1997) – although I do not defend the form in which it is generally practiced. That predilection obviously colors my reactions to the present articles.

[3] As an aside and as not accurately conveyed by McLean and Ernest, we (Levin & Robinson, in press; Robinson & Levin, 1997) do not argue *that* alternative language is needed in Results sections. Rather, we suggest that *if* better language is mandated, then descriptors such as "statistically real," "statistically nonchance," and "statistically different" could readily say what one means and mean what one says without a trace of "significance." [4] A primary function of statistical hypothesis testing has been analogized in even more colorful terms – a "crap detector" – by a distinguished scholar who shall unfortunately remain nameless in that I cannot locate the appropriate citation at the moment.

[5] In each case, the obtained treatment difference is statistically "real," or nonchance (*p* # .05, one-tailed), on the basis of either a parametric or nonparametric hypothesis test.

[6] The major problem in this example arises from the conditions' differing variabilities. That problem could be accounted for by defining alternative *d*-like effect-size measures based on just the control condition's (Condition A's) standard deviation, as has been suggested by Glass (1977), Hedges and Olkin (1985), and others. Interpreting effect sizes, in the absence of raw data, remains a problem for ç^{2} and Cohen's *d*, however. Concerns about effect sizes based on relative metrics, and a variety of other concerns, are detailed by O'Grady (1982), Frick (1995), and Fern and Monroe (1996).

[7] Note that assumptions violations also affect the validity of other inferential statistical alternatives, such as confidence intervals and meta-analyses. Interestingly and in contrast to the "replication" objectives misattributed to them, bootstrapping and jackknifing are methods that *do* possess either "distribution-free" or other robust qualities that could be exploited to circumvent assumption- violations problems.

[8] In this example, I have tried to mitigate the important "base-rate" problem (e.g., Derry, Levin, & Schauble, 1995) by restricting the population to patients with an "at risk" profile. Even so, the problem remains and would need to be taken into account should the screening test's results prove positive.

[9] On the other hand, if it can be documented that the major impediment to scientific progress lies in the value-lessness of reporting single- or few-study investigations (as some have accused), then why not simply discontinue the production of journals that publish primary-research articles and continue with only those that publish research syntheses? Imagine what a triumph that would be for meta-analysis enthusiasts!