Reporting Practices in Quantitative Teacher Education Research: One Look at the Evidence Cited in the AERA Panel Report

Linda Reichwein Zientek, Mary Margaret Capraro, and Robert M. Capraro

The authors of this article examine the analytic and reporting features of research articles cited in *Studying Teacher Education: The Report of the AERA Panel on Research and Teacher Education* (Cochran-Smith & Zeichner, 2005b) that used quantitative reporting practices. Their purpose was to help to identify reporting practices that can be improved to further the creation of the best possible evidence base for teacher education. Their findings indicate that many study reports lack (a) effect sizes, (b) confidence intervals, and (c) reliability and validity coefficients. One possible solution is for journal editors to emphasize clearly the expectations established in *Standards for Reporting on Empirical Social Science Research in AERA Publications* (AERA, 2006).

Keywords: empirical research; evidence-based practice; research methodology; teacher education

Education research yields its greatest potential contribution to an evidence base for practice when recommended reporting practices are followed and balanced with flexibility and creativity (American Educational Research Association [AERA], 2006; Natriello, 2000; Wilkinson & Task Force on Statistical Inference [TFSI], 1999; Zeichner, 2005). As testimony to the importance of teacher education research and corresponding research practices, AERA recently published two major reports: a panel study synthesis of research titled *Studying Teacher Education: The Report of the AERA Panel on Research and Teacher Education* (Cochran-Smith & Zeichner, 2005b); and a set of AERA standards entitled *Standards for Reporting on Empirical Social Science Research in AERA Publications* (AERA, 2006). Both of these reports provide assistance to editors, reviewers, and education researchers charged with vetting new research and improving the overall quality of education research.

In the present article, we examine the analytic and reporting features of research studies cited in *Studying Teacher Education* (Cochran-Smith & Zeichner, 2005b) that used quantitative reporting practices. Our purpose is to help identify quantitative reporting practices that can be improved in the interest of creating the best possible evidence base for teacher education.

Reporting Practices and Standards

Reporting practices are important in communicating findings. In teacher education, we hope that our findings ultimately will precipitate improvements in the classroom. Stipek (2005) noted that "both the desire to consult research and the skills to interpret it will need to be developed within the teaching community" (p. 33); however, teachers are often not the ones reading the research. Therefore, we will need to educate those who prepare teachers and those who make important decisions within schools (Wolk, 2007). A prerequisite is that published research include necessary and sufficient information to enable readers to make well-informed evaluations of the warrants for research conclusions. Archival publications provide a record of research methods and findings, a historical perspective on influences on education, and a foundation for future research.

Standards facilitate research quality by articulating a set of common expectations for editors, editorial boards, reviewers, and researchers. According to AERA (2006), the specifications of reporting standards are useful for researchers as they prepare manuscripts, for editors and reviewers as they review manuscripts, and for readers as they attempt to build their practice or store of knowledge on the basis of the published works. Feuer, Towne, and Shavelson (2002) state that

> researchers must have a clear, commonly held understanding of how scientific claims are warranted. . . . [I]t is incumbent upon the field to cultivate its own form of life including, however difficult this may be, attention to bolstering research quality. (p. 9)

Aspects of research. Ernest and McLean (1998) reason that conducting research is akin to detective work and that as researchers we need to provide readers with as many clues as reasonably possible regarding how we have reached our conclusions. Researchers need to provide enough information to allow readers to formulate their own well-informed interpretations of results, and enough information to allow meta-analysts to evaluate studies across samples and time. The ideas of *warrant* and *transparency* are two overarching principles of the AERA (2006) Standards, which assert that "adequate evidence should be provided to

Educational Researcher, Vol. 37, No. 4, pp. 208–216
DOI: 10.3102/0013189X08319762

justify the results and conclusions" and that the "logic of inquiry and activities that led from the development of the initial interest, topic, problem, or research question . . . to the articulated outcomes of the study" should be made evident (p. 33).

Our research questions should be clear and should drive our selection of research methods (Raudenbush, 2005). The analyses that are conducted should be designed "to shed light on theory" (Pedhazur, 1997, p. 8). According to Thompson (2006), "Good social science research is primarily about thinking, about reflection, and about judgment" (p. v). We should reflect on the methods we use. Rather than simply focus on the technicalities of statistical procedures, we should "learn when to use them" (Quilici & Mayer, 1996, p. 144). Likewise, Wilkinson and TFSI (1999) criticize the "thoughtless application of statistical methods" (p. 603).

For readers to make well-informed decisions on the basis of a study, the essential elements of the study's design, sample, and analyses should be reported. The design of a study should be clear. Descriptions of the sampling procedures should explain the rationale for the inclusion and exclusion criteria and the sample size. When possible, researchers should define characteristics of the population and compare the sample characteristics with those of the population (AERA, 2006). Score validity and reliability should be investigated for data in hand (M. M. Capraro, Capraro, & Henson, 2001; R. M. Capraro & Capraro, 2002; Thompson, 2003). Statistical analyses should be clear, and effect sizes and confidence intervals should be reported (AERA, 2006). Both the AERA (2006) Standards and the AERA Panel on Research and Teacher Education (Cochran-Smith & Zeichner, 2005b) recommend the inclusion of sufficient information to allow replication.

Peer-review process. The peer-review process was established to advantage the highest quality research in obtaining access to scarce journal space. Kassirer and Campion (1994), Pedhazur (1997), and Sumner et al. (2000) addressed the need to improve the peer-review process. Recommendations include ensuring that reviewers are educated about analytic methods and that editors require high-quality reviews (M. M. Capraro & Capraro, 2003). Pedhazur emphasized the message that educating reviewers is crucial. As stated by Pedhazur, "Detection of many egregious errors requires nothing more than careful reading," and in order "to detect errors in the application of an analytic method, the reviewer ought to be familiar with it" (p. 13). Zeichner (2005) also addressed the need to offer guidance to reviewers.

Modern statistical reform. In 1999, Wilkinson and the TFSI recommended certain reporting practices, which were reiterated in the *Publication Manual of the American Psychological Association* (APA; 2001) and in the AERA (2006) Standards. These recommendations include reporting inferential statistics, means, standard deviations, *p* values, effect sizes, and confidence intervals (AERA, 2006; Thompson, 2002, 2006; Wilkinson & TFSI, 1999).

Prior to the publication of the recommendations of Wilkinson and the TFSI (1999), there was a lengthy discussion on the importance of effect sizes and a debate on the use of null hypothesis statistical significance testing (NHSST). This debate occurred in diverse disciplines, including education (e.g., Thompson, 1996), psychology (cf. Cohen, 1994; Kirk, 1996; Schmidt, 1996), economics (Ziliak & McCloskey, 2004), and the life sciences (see Anderson, Burnham, & Thompson, 2000). One result was the 1996 formation of the APA Task Force to address issues related to statistical significance testing and effect size reporting.

In 2001, the fifth edition of the APA *Publication Manual* listed "failure to report effect sizes" as one of the "defects" that editors and reviewers watch for in evaluating a research paper (p. 5). In 2006, AERA adopted its publication Standards, which require reporting of effect sizes and encourage the use of confidence intervals as well. Today, more than two dozen journals require effect size reporting, as stated in their author guidelines.

Methodology

Article Selection

The AERA Panel on Research and Teacher Education, whose task was to provide a critical analysis of the "empirical evidence relevant to practices and policies in preservice teacher education" (Cochran-Smith & Zeichner, 2005a, p. 1), established criteria for the inclusion and exclusion of studies. In the present study we limited our consideration to cited articles on specific topic areas of teacher education research. The topics included pathways to teaching, novice teacher characteristics, subject-matter and field-based teacher preparation, pedagogical approaches to teacher preparation, preparing teachers for special-needs students, teacher quality indicators, and accountability in teacher education. Some chapters cited more research articles because of the "history, scope, and depth" (p. 4) of available research related to their topics.

We examined articles cited in the 2005 AERA report (Cochran-Smith & Zeichner, 2005b) containing quantitative analytic methods. If the analytic method could not be ascertained from the article abstract, we obtained the article to determine suitability. Books, qualitative studies, studies identified as mixed methods, and meta-analytic studies were excluded, resulting in 174 articles that were read and coded. Twenty-two of these included both quantitative and qualitative data but were not classified as mixed methods. Of the 174 articles, 36 reported descriptive statistics or percentages but did not employ hypothesis testing.

Coding

Our coding scheme was based on recommendations contained in the AERA (2006) Standards and in Wilkinson and TFSI (1999). We coded reporting practices for the following major categories: (a) demographics, (b) comparisons of the sample with the population, (c) use of internal and external replicability techniques, (d) means and standard deviations, (e) statistical significance, (f) score validity and reliability descriptions, (g) confidence intervals, (h) statistical analyses, (i) inferential statistics, (j) effect sizes, and (k) pictorial representations. Each category contained numerous indicators that were scored dichotomously in each article as either one or zero (1 = *present,* 0 = *not present*). For statistical analyses, if a test was conducted, the scoring sheet was marked 1 for that category (e.g., analysis of variance [ANOVA] or multivariate analysis of variance [MANOVA]). When a statistical analysis was conducted, the reporting practices for *p* values, inferential statistics, and effect sizes were also dichotomously coded (see Appendix for coding scheme).

Interrater Reliability

Because of the large number of articles coded, interrater reliability for the three authors/raters was estimated to ensure "like mindedness." To quantify the level of consistency among the raters, an index of interrater reliability was determined (Huck, 2004). First, each of the three researchers independently coded five randomly chosen articles. Discussions and reconciliations were held for instances where like mindedness was not achieved. At that time, all discrepancies were resolved concerning the purposes and uses of all categories. Next, another five articles were chosen at random and coded individually by each of the three researchers. Results for the interrater reliability were just over 95%. Finally, the articles were divided among the three researchers, and all of the remaining articles were coded. In instances where a researcher had a question concerning an article, a discussion was held. All discussions and reconciliations were completed before the quantitative analysis was conducted. This process was similar to that recommended by Beatty (2004).

To prevent rater drift and to minimize the effects of repeated coding, intrarater reliability was also estimated. Intrarater reliability was estimated individually for each researcher by randomly selecting from that rater's previously coded articles and the other raters recoding these articles. The consistency estimates for the raters were 0.99, 0.98, and 0.99, respectively.

Results

Replicability

During the 1990s, an increasing number of researchers recognized, based on the arguments of Cohen (1994), Thompson (1996), and others, that NHSST does not evaluate result replicability. Cumming (in press) recently published a seminal Monte Carlo study on these dynamics. Consequently, because finding results that can be replicated under stated conditions is important in quantitative research, scholars began emphasizing the use of methods that evaluate result replicability.

Internal replication. The majority of the studies we selected did not employ internal replication methods, such as cross-validation, the jackknife, or the bootstrap. Four studies reported internal replicability analyses; one used a cross-validation, one the bootstrap, and two the jackknife.

Potential external replication. At a minimum, if a study is to be replicated, sufficient information regarding sample selection, demographics, survey instruments, and methods of analysis must be provided (Cronbach, 1982). In addition, for enhanced replication, sufficient details need to be provided on how the researchers obtained "access and ethical approval . . . to undertake the study" and on "recruitment of participants; how they gave consent; where the study was undertaken," and "the development of data collection instruments" (Thomson, 2004, p. 297).

In total, 9% ($n = 15$) of the articles included all of the elements necessary to possibly conduct a replication study. The following criteria were coded for potential external replicability: (a) Participant selection was explicitly described, (b) demographics were provided (i.e., at least ethnicity or gender), (c) the sample was directly compared with the population, (d) the instrument was available when an instrument was used, and (e) techniques and procedures were apparent.

The majority of the studies (98%) reported sample sizes. Participant selection was clearly explained in 83% ($n = 144$) of the studies. Our analysis of reporting practices on demographics was limited to studies that consisted of samples of people ($n = 172$). In 60% ($n = 103$) of the studies, one or more of the following were reported: ethnicity, gender, or age. In 45% ($n = 77$) of the studies, ethnicity was reported. In 52% ($n = 89$) of the studies, gender was reported. Furthermore, age was reported in 38% ($n = 66$) of the studies. Of the 174 studies, 17% ($n = 29$) explicitly compared the sample characteristics with those of the population on ethnicity (5%), age (3%), gender (4%), and other characteristics (13%). Some studies compared on multiple characteristics.

Eighty-three percent ($n = 144$) of the studies based at least a portion of their findings on surveys. Of the 144 surveys, 110 were not based on national survey findings. Forty-eight percent of those surveys ($n = 53$) were available in the article or from the author or other sources. When the results were based on national survey findings ($n = 34$), most of these surveys were available on the Internet or through the national agencies that conducted them. The majority (98%) of reports identified their testing procedures.

Means, Standard Deviations, and Inferential Statistics

Sixty-five percent of the studies ($n = 113$) reported means, and 44% ($n = 77$) reported both means and standard deviations. As Thompson (2006) emphasized,

> Because central tendency statistics more accurately represent all the scores as dispersion is smaller, dispersion statistics (e.g., the *SD*) characterize how well central tendency statistics do at representing the data. Therefore, *always* report the *SD* whenever reporting the mean. (p. 72)

Seventy-nine percent ($n = 138$) of the studies conducted statistical analyses. Multiple NHSST tests were often conducted within studies. Table 1 presents the tests conducted with the frequency of inclusion of the obtained inferential test statistic. Regressions, ANOVAs, and *t* tests were the predominant analyses used by education researchers.

Of the seven MANOVAs, one was not statistically significant; five were followed by post hoc ANOVAs; and one preceded a MANOVA with ANOVAs. In the last case, the statistically significant results of the ANOVA were used to justify the use of a MANOVA. Inferential statistics were reported for 76% of the univariate tests, 78% of the multivariate tests, 96% of the bivariate correlation tests, 88% of the regressions, 50% of the other types of parametric tests, and 82% of the nonparametric tests.

Exact p *Values*

Of the 138 studies that used statistical analyses, 38% ($n = 52$) reported exact *p* values and 29 reported exact *p* values for statistically nonsignificant results. Some studies reported statistical significance without providing the exact *p* value. The simple error of incorrectly rounding *p* values to zero or less than zero resulted in the reporting of *p* values as zero in three articles and as less than zero in one article, both of which are impossible.

Table 1
Reporting of Inferential Statistics and Effect Sizes by Type of Test Conducted

Test Type	*n*	Reported Inferential Statistic (*n*)		Reported Effect Size (%)	
		Yes	%	Yes	%
Correlations	28	27	96	27	96
Univariate					
t test	36	25	69	3	8
Paired *t* test	8	8	100	0	0
ANOVA	34	26	76	2	6
ANCOVA	8	6	75	1	13
Log-linear analysis	0	–		–	
Other parametric[a]	8	4	50	1	13
Nonparametric					
Chi-square	25	21	84	0	0
Other nonparametric	9	7	78	1	11
Regressions					
OLS/WLS	8	8	100	8	100
Multiple	17	13	76	16	94
Stepwise	4	3	75	3	75
Logistic	4	4	100	4	100
Multilogistic	1	1	100	1	100
HLM	9	9	100	8	89
Multivariate					
MANOVA/ MANCOVA/ Hotelling's T^2	10	9	90	1	10
CCA	2	1[b]	50	1	50
DDA/PDA	0	–		–	
Other multivariate[c]	6	4	67	3	50
Total	217	176	81	80	37

Note. Multiple tests were conducted within studies. OLS = ordinary least squares; WLS = weighted least squares; HLM = hierarchical linear modeling; CCA = canonical correlation analysis; DDA = descriptive discriminant analysis; PDA = predictive discriminant analysis.

[a]The "other parametric" category consisted of the following: three unidentified tests, one Fisher's LSD test, one Scheffe test, one *z* test, and two probit models.

[b]One researcher discussed having conducted a CCA on the data but elaborated on why results from another analysis were reported.

[c]The "other multivariate" category consisted of one Wald's test of significance, one multivariate test, two Cox proportional hazard models, and two SEMs.

Fifty-seven percent ($n = 78$) of the 138 studies referred to the null-hypothesis test results with an a priori alpha as being "statistically significant," as recommended by Thompson (1996); others (43%) simply used the term "significant," which may more easily be misinterpreted as meaning "important." However, further investigation indicated that in recent years researchers were doing far better. From 1997 onward, the majority (72%) wrote that their results were "statistically significant," as compared with 44% before 1997.

Confidence Intervals and Effect Sizes

A scant 4% of the articles ($n = 7$) reported confidence intervals. For better or worse, however, this dismally small figure is fully consistent with the usage reported elsewhere in educational and psychological research (Finch, Cumming, & Thomason, 2001; Kieffer, Reese, & Thompson, 2001).

Thirty-nine percent of the cited articles ($n = 68$) reported effect sizes. However, a smaller number (i.e., 45 of these 68) also interpreted the reported effect sizes. Table 1 shows the testing procedures and reporting practices of researchers by statistical test. If the Pearson or Spearman rho correlations were conducted, effect sizes were reported 96% of the time. The majority of researchers reported effect sizes for regressions; however, researchers were inconsistent in their reporting of effect sizes for other univariate, multivariate, and nonparametric tests. Effect sizes were reported for 7% of univariate tests, 28% of multivariate tests, 96% of correlations, 93% of regressions, 13% of other parametric tests, and 3% of nonparametric tests. Sixty-seven percent of the articles reporting effect sizes interpreted them; however, interpretation was limited mostly to regressions and correlations.

Validity and Reliability

There are several types of validity, including content, concurrent, construct, and predictive (Crocker & Algina, 1986). The most commonly reported is construct validity. Several methods can be used to examine construct validity, including correlations between measures of the construct, differentiation between groups, factor analysis, multitrait-multimethod matrix (Crocker & Algina, 1986), and structural equation modeling (SEM; Jöreskog, 1977; Kline, 2005). In the present study, the use of SEM and factor analysis was coded.

One percent ($n = 2$) of the cited articles conducted an SEM. Six percent ($n = 11$) of the cited articles conducted a factor analysis on the data in hand, with five reporting the rotation methods and three reporting both the pattern and structure coefficients. Thirteen percent ($n = 22$) reported reliability for their data; nine of those incorrectly attributed the reliability to the instrument. As Wilkinson and the TFSI (1999), Thompson (2003), M. M. Capraro et al. (2001), and others have emphasized, tests are *not* reliable; scores are, and reliability should routinely be reported for the actual data being analyzed in a given study.

In addition, four studies used reliability induction where they reported reliability estimates from previous studies without explicitly comparing sample compositions from the present with the prior studies. Three of these four incorrectly reported the prior estimates of score reliability!

Pictorial Representations

Thirty-three percent ($n = 58$) of the articles contained figures. The figures took a great variety of forms, including a box-and-whisker plot (1), regression/trend lines (6), bar graphs or histograms (30), line graphs (13), and other figures (28).

Discussion

There is a circular dependency among research, policy, and doctoral research training. Policy drives research, creating an ebb and flow between uses of descriptive and causal designs in a given policy arena. This ebb and flow between research designs, in turn, creates a continuum of research foci across a range of courses and requirements for education researchers (R. M. Capraro & Thompson, in press). The cycle becomes self-propagating because those trained

during any given phase persist until the newly trained, who are educated under new policy, assume the mantle of leadership (R. M. Capraro & Thompson, in press; Natriello, 2000). How researchers are trained is influenced by policy; policy affects the research they conduct; and subsequently, the research that is published influences the policies that are made. For this reason, researchers must provide sufficient information to enable educators to make well-informed decisions, and Ph.D. programs must adequately prepare competent researchers to conduct empirical research.

Adopting standards is one method of guiding research practices and facilitating the production of high-quality research that can be held up to scrutiny, used as a basis for future research, and, when appropriate, replicated. Standards can also form the policy base for the training and preparation of future researchers (AERA, 2006; Aiken, West, & Millsap, 2008; Aiken, West, Sechrest, & Reno, 1990; Henson & Williams, 2006). Over the past decade, a consensus on reporting practices has emerged (AERA, 2006; Wilkinson & TFSI, 1999). In the present study, we evaluated research reporting practices in vetted teacher education literature and sought to identify deficiencies that need to be addressed in future work. These results will serve as a benchmark to determine the impact of new standards on the literature in our field.

Sample Description

One *source of evidence* recommended in the AERA (2006) Standards is a description of the *unit of study* that includes, but is not limited to, the number of participants and the characteristics of the sample. Most studies reported sample size and sample selection procedures. Demographics were provided in just over half of the studies. Because the teacher population historically has been and continues to be predominantly White and female, some researchers may have felt that reporting demographics would not add valuable information to their articles. We argue that, regardless of the lack of diversity in samples, sample demographics should be reported because they enable readers to evaluate representativeness and to understand which studies have similar samples or describe similar populations.

One method of determining sample representativeness is to compare explicitly the sample with the population. In the present set of studies, the majority (83%) did not compare sample characteristics with those of their populations. This omission, like the omission of demographics in general (40% of the studies did not report demographics), limits the ability of readers to determine representativeness and limits the ability of researchers to conduct meaningful meta-analyses.

Score Validities and Reliabilities

Another source of evidence recommended in the AERA (2006) Standards is reporting on data collection. In the present set of studies, surveys were the dominant mode for collecting data, but only about half of the researcher-developed surveys were made available (national surveys were usually available through the researching agency). If a study is to be externally replicated, the instrument(s) must be available. Zeichner (2005) noted that "many studies reviewed in [*Studying Teacher Education*] provide no information about how instruments used for data collection were developed" (p. 741). Zeichner further noted that, when researchers develop and administer instruments but do not provide the instruments and then fail to provide validity and reliability estimates for the data in hand, this lack of information "necessarily weakens the claims that researchers can make about the effects of what was examined" (p. 741).

Multiple approaches exist for testing construct validity (e.g., correlations between measures of the construct, differentiation between groups, factor analysis, multitrait-multimethod matrix, SEM); however, in the present study, we limited our validity coding to SEM and factor analysis (Crocker & Algina, 1986; Kline, 2005). As noted by Henson and Roberts (2006), "Thanks to the advent of technology, factor analysis is now frequently employed in both measurement and substantive research" (p. 394). A potential danger is the ease with which factor analysis results can be produced by misguided novice researchers who do not understand that a factor analysis consists of more than using the default options in a statistical package but, rather, "involves a linear sequence of decisions each involving a menu of several available choices" (Thompson, 2004, p. 27). To enable readers to determine the appropriateness of the methods used and to evaluate the findings, factor analysis methods and results need to be reported in detail. In the present set of articles, validity estimates were, for the most part, not reported. The inadequate use of SEM and factor analysis suggests that Ph.D. graduates entering the research field may not be sufficiently competent to conduct confirmatory factor analysis, exploratory factor analysis, or SEM (Henson & Williams, 2006).

In general, reliability coefficients also were not reported for the data in hand, and almost half of the studies that did report reliability estimates for their data or from previous studies incorrectly described reliability as being a function of the test. Reliability is inherently related to the sample and therefore should always be reported for the data in hand (cf. Thompson, 2003; Wilkinson & TFSI, 1999). Reliability affects the ability to detect both statistical significance and noteworthy effect sizes. Failure to report reliability coefficients may lead to misinterpretations, and studies may be conducted that cannot produce noteworthy effect sizes (Thompson, 1994, 2003). According to Reinhardt (1996),

> Reliability is critical in detecting effects in substantive research. For example, if a dependent variable is measured such that scores are perfectly unreliable, the effect size in the study will unavoidably be zero, and the results will not be statistically significant at any sample size, including an incredibly large one. (p. 3)

Effect Sizes

According to Thompson (2000), "All parametric statistical analyses are special cases within a single general linear model (GLM) family" (p. 262), and all methods are correlational in nature and yield variance-accounted-for effect size estimates. Effect sizes should be reported for both statistically significant and statistically nonsignificant results (AERA, 2006; Wilkinson & TFSI, 1999). The majority of the studies conducted univariate or nonparametric tests without including effect sizes.

Not reporting effects can be detrimental. An infinitesimal p value does not make the effect magnitude important. And for large enough sample sizes, even diminutive effect sizes will be statistically significant and the nil null will be rejected (R. M. Capraro, 2006;

Thompson, 2006, 2007). Two studies with different sample sizes can have the same effect size with one indicating statistically significant results and the other having statistically nonsignificant results. Therefore, a requisite power analysis should be conducted before beginning a study. The issue is that without sufficient sample size, given an anticipated effect, one cannot be sure whether the statistically nonsignificant result is just an artifact of sample size (Cohen, 1992).

Reporting effect sizes contextualizes the impact of the treatment (Thompson, 2007). In the present review of research, except for cases of regressions and correlations, researchers infrequently reported effect sizes. Our results corroborated Kirk's (1996) findings that researchers most frequently reported effect sizes commonly produced by statistical packages such as "variance-accounted-for" effect size estimates and R^2. Hence we cannot discern whether researchers conducting regressions typically understand effect sizes or are merely reporting the information routinely provided by the statistical package.

Means and Standard Deviations

In addition to reporting effect sizes, both means and standard deviations should be reported. Reporting descriptive statistics is advocated by the AERA (2006) Standards. Although 65% of the present set of studies reported means, a third of those did not report *both* means and standard deviations. Therefore, it was not evident that the researchers understood the value of quantifying the quality of the mean by characterizing the score spread of the data.

Statistical Procedures

Almost all of the researchers identified the statistical analyses that they had conducted (98%), and multiple NHSST analyses were often conducted within a single study. Statistical procedures are important, but we do not want to neglect the importance of (a) knowing when to use the appropriate procedures (Quilici & Mayer, 1996) and (b) the theoretical framework determining the variables and the appropriate analytic techniques (Pedhazur, 1997). As recommended in the AERA (2006) Standards,

> Reporting should clearly state *what statistical analyses were conducted and the appropriateness of the statistical tests,* linking them to the logic of design and analysis and describing them in enough detail that they could be replicated by a competent data analyst. (p. 37)

In the present study, findings presented in Table 1 show that ANOVAs, *t* tests, and regressions were the most commonly conducted tests. These results are similar to Kirk's (1996) review of literature, and they corroborate findings that Ph.D. graduates have been entering the field more comfortable using regressions and ANOVAs than any other statistical methods (Aiken et al., 2008; Aiken et al., 1990; Henson & Williams, 2006). Statistically significant MANOVAs should be followed by a multivariate post hoc method, such as the use of descriptive discriminant analysis (Enders, 2003; Huberty, 1994) and the Roy-Bose simultaneous confidence intervals (Stevens, 2002). Some researchers have advocated post hoc ANOVAs with Bonferroni corrections following statistically significant MANOVAs (Stevens, 2002); however, if the nature of the problem is multivariate, then using univariate post hoc analyses to explore results yields noncommensurate results that are logically apples to oranges (Enders, 2003). Eighty-three percent of the statistically significant MANOVAs were followed by a post hoc univariate (ANOVA) analysis. Thus the findings in the present study support Enders's statement that, "despite the clear message from the methodological literature, multiple univariate tests are still the predominant method for examining group differences following a significant MANOVA" (p. 41).

Exact p *values.* Fewer than half (38%) of the studies for which the authors had conducted statistical analyses reported exact *p* values, although such reporting is required by the APA (2001) *Publication Manual.* In addition, three researchers incorrectly reported *p* as zero, and one incorrectly reported *p* as less than .000 (researchers should not round *p* values to zero and then report *p* values as being equal to—or less than—zero).

Statistically significant. A change in reporting practices occurred after 1996, suggesting that discussions about those practices did affect subsequent published literature (cf. Cohen, 1990, 1994; Kirk, 1996; Schmidt, 1996; Thompson, 1996; Wilkinson & TFSI, 1999). Differentiating between the terms *significant* and *statistically significant* is important. The latter indicates that a null hypothesis test was conducted and that $p_{calculated}$ was less than the a priori alpha level. In contrast, the term *significant* wrongly implies to at least some readers that results are noteworthy or replicable. Therefore, the language we use should be precise and avoid ambiguity so that a wider audience can more easily access our evidence base. Because statistical significance testing is dependent on sample size and does not evaluate result importance or replicability, researchers conducting null-hypothesis testing should be reporting confidence intervals (AERA, 2006; Thompson, 1996).

Confidence Intervals

Confidence intervals are very important tools of results description, measuring "how sure" we are of our results (M. M. Capraro, 2005). Yet only 4% of the cited articles reported confidence intervals. Some researchers think that confidence intervals are just another method for testing statistical significance (Knapp & Sawilowsky, 2001). A difference between confidence intervals and NHSST lies in the fact that confidence intervals can be constructed even without formulating a null hypothesis (Thompson, 2006). As stated by Wilkinson and the TFSI (1999), "*It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual* p *value or, better still, a confidence interval*" (p. 599). Confidence intervals across studies tell us how accurately and consistently effects operate over time. The APA Task Force suggested that confidence intervals should *always* be reported (Wilkinson & TFSI, 1999), and the APA *Publication Manual* (2001) called confidence intervals "the best reporting strategy" (p. 22).

Pictorial Representations

Another way to provide sufficient information is to include pictorial representations of results (Wilkinson & TFSI, 1999). Thirty-three percent of the articles contained figures, but these were limited mainly to bar graphs, histograms, or circle graphs.

Box-and-whisker plots were reported in only one study. Visual representations are important because they can aid in the detection of trends within data and help identify outliers and score clusters (Billstein, Libeskind, & Lott, 2007; Hinkle, Wiersma, & Jurs, 1998; Thompson, 2006).

The development of technology has facilitated our ability to graphically inspect and report data, and graphics are relatively easy to create with modern statistical software packages. Although at least a third of the researchers understood the value of visual representations, the fact that only one used the box-and-whisker plots suggests that people may not understand the amount of information that such plots can display. Box-and-whisker plots display (a) dispersion of scores, (b) median values, and (c) outliers, all in one accessible figure. Hence researchers can easily compare groups, view the spread of data, and identify the impact of outliers on results.

Conclusions

To enable educators to make well-informed decisions, researchers have the responsibility of ensuring that methodologies are appropriate, that results and conclusions are justified (i.e., warranted), and that the development, implementation, and analysis leading to the research findings are clear (i.e., transparent). Education research has made progress, but there is still room for improvement in the quantitative literature in teacher education. Adhering to the two overarching principles on which the AERA (2006) Standards were based—"the sufficiency of the warrants and the transparency of the report" (p. 33)—will further "AERA's broader educational mission to advance high-quality research in education and to foster excellence in reporting on empirical research" (p. 33).

The findings from the present study indicate that (a) effect sizes, (b) confidence intervals, and (c) reliability and validity coefficients are lacking from many study reports. The current review of literature supports the notion that deficits in doctoral preparation are reflected in research practices. Areas of methodology that were lacking in graduate programs identified by Aiken et al. (2008), R. M. Capraro and Thompson (in press), and Henson and Williams (2006) mean that some researchers may not be fully trained to conduct research that honors the precepts of contemporary standards. One possible solution is for journal editors to emphasize clearly these three expectations for reports in AERA publications (AERA, 2006). With the recognition that effect sizes are important, we can hope that the inclusion of effect sizes will finally become a standard practice in future publications.

Researchers also need to understand the importance of providing instruments, comparing characteristics of the sample with the population, and investigating result replicability. In accordance with research recommendations, researchers ought to differentiate between "significant" and "statistically significant" results by (a) using the word *significant* only to refer to NHSST results and then (b) always using both words in the phrase *statistically significant*.

As noted by Zeichner (2005), despite the problems with the extant evidence base that have been identified, "there is reason to be optimistic about the future" (p. 755). Recommendations for improvements include providing "more detailed guidance to reviewers of research in the major peer-reviewed journals in teacher education" (p. 756). In the present study, we have sought to aid in addressing areas of statistical reporting practices that can be improved. The implementation of the AERA (2006) Standards and an awareness of contemporary analytic expectations will improve education research and ensure that high-quality research is produced and published. Our findings suggest that schools of education and professional organizations together need to promote awareness among education researchers regarding the importance of effective reporting practices. Our findings can also serve as a benchmark for determining the future impact of the recent standards on research reporting over the long term.

REFERENCES

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63,* 32–50.

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45,* 721–734.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management, 64,* 912–923.

Beatty, P. (2004). The dynamics of cognitive interviewing. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, et al. (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 45–66). New York: Wiley.

Billstein, R., Libeskind, S., & Lott, J. W. (2007). *A problem solving approach to mathematics for elementary school teachers* (9th ed.). Boston: Addison-Wesley.

Capraro, M. M. (2005). An introduction to confidence intervals for both statistical estimates and effect sizes. *Research in the Schools, 12*(2), 22–32.

Capraro, M. M., & Capraro, R. M. (2003). Exploring the APA fifth edition publication manual's impact on the analytic preferences of journal editorial board members. *Educational and Psychological Measurement, 63,* 554–565.

Capraro, M. M., Capraro, R. M., & Henson, R. K. (2001). Measurement error of scores on the Mathematics Anxiety Rating Scale across studies. *Educational and Psychological Measurement, 61,* 373–386.

Capraro, R. M. (2006). Significance level. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics.* Thousand Oaks, CA: Sage.

Capraro, R. M., & Capraro, M. M. (2002). Myers-Briggs Type Indicator score reliability across studies: A meta-analytic reliability generalization study. *Educational and Psychological Measurement, 62,* 590–602.

Capraro, R. M., & Thompson, B. (in press). What is an educational researcher, and what kinds of research will the educational researchers of tomorrow be trained to do? *Journal of Educational Research.*

Cochran-Smith, M., & Zeichner, K. M. (2005a). Executive summary. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA Panel on Research and Teacher Education* (pp. 1–36). Mahwah, NJ: Lawrence Erlbaum.

Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2005b). *Studying teacher education: The report of the AERA Panel on Research and Teacher Education*. Mahwah, NJ: Lawrence Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.
Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.
Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth.
Cronbach, L. J. (1982). *Designing evaluations of educational and social programs.* San Francisco: Jossey-Bass.
Cumming, G. (in press). Replication and *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science.*
Enders, C. K. (2003). Performing multivariate group comparisons following a statistically significant MANOVA. *Measurement and Evaluation in Counseling and Development, 36,* 40–56.
Ernest, J. M., & McLean, J. E. (1998). Fight the good fight: A response to Thompson, Knapp, and Levin. *Research in the Schools, 5*(2), 59–62.
Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31*(8), 4–14.
Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology:* Little evidence of reform. *Educational and Psychological Measurement, 61,* 181–210.
Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66,* 393–416.
Henson, R. K., & Williams, C. (2006, April). *Doctoral training in research methodology: A national survey of education-related degrees.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences* (4th ed.). Boston: Houghton Mifflin.
Huberty, C. (1994). *Applied discriminant analysis.* New York: Wiley.
Huck, S. (2004). *Reading statistics and research* (4th ed.). Boston: Pearson.
Jöreskog, K. G. (1977). Structural equation models in the social sciences: Specification, estimation, and testing. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 265–287). Amsterdam: North Holland.
Kassirer, J. P., & Campion, E. W. (1994). Peer review: Crude and understudied, but indispensable. *Journal of the American Medical Association, 272,* 96–97.
Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *Journal of Experimental Education, 69,* 280–309.
Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.
Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.
Knapp, T., & Sawilowsky, S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education, 70,* 65–79.
Natriello, G. (2000). Research and educational policy (Part 1): The relationship between research and policy. *Teachers College Record.* Retrieved July 1, 2007, from http://www.tcrecord.org, ID No. 10636
Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.
Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88,* 144–161.
Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34*(5), 25–31.
Reinhardt, B. M. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 3–20). Greenwich, CT: JAI.
Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.
Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.
Stipek, D. (2005). Scientifically based practice: It's about more than improving the quality of research. *Education Week, 24*(28), 33, 44.
Sumner, T. R., Shum, S. B., Wright, M. J., Bonnardel, N., Piolat, A., & Chevalier, A. (2000). Redesigning the peer review process: A developmental theory-in-action. In R. Dieng, A. Giboin, L. Karsenty, & G. De Michels (Eds.), *Designing cooperative systems.* Amsterdam, The Netherlands: IOS Press.
Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality, 62,* 157–176.
Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30.
Thompson, B. (2000). Ten commandments of structural equation modeling. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261–284). Washington, DC: American Psychological Association.
Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25–32.
Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.
Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.
Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44,* 423–432.
Thomson, A. (2004). What should be included in a methods section? *Midwifery, 20,* 297–298.
Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.
Wolk, R. A. (2007). Education research could improve schools, but probably won't. *Education Week, 26*(42), 38–39.
Zeichner, K. M. (2005). A research agenda for teacher education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA panel on research and teacher education* (pp. 737–759). Mahwah, NJ: Lawrence Erlbaum.
Ziliak, S. T., & McCloskey, D. N. (2004). Size matters: The standard error regressions in the *American Economic Review. Journal of Socio-Economics, 33,* 527–546.

AUTHORS

LINDA REICHWEIN ZIENTEK is an assistant professor of mathematics education at Sam Houston State University, Department of Mathematics and Statistics, P.O. Box 2206, Huntsville, TX 77341; *lrzientek@shsu.edu.* Her research focuses on mathematics teacher preparation, teacher induction programs, and quantitative research methods.

MARY MARGARET CAPRARO is an assistant professor of mathematics education at Texas A&M University, Department of Teaching,

Learning and Culture, Mail Stop 4232, College Station, TX 77843; *mmcapraro@tamu.edu.* Her research focuses on effective teaching practices and mathematics teacher education preparation.

ROBERT M. CAPRARO is an associate professor of mathematics education at Texas A&M University, Department of Teaching, Learning and Culture, Mail Stop 4232, College Station, TX 77843; *rcapraro@coe.tamu.edu.* His research focuses on mathematical representation, teacher knowledge, and assessment.

Manuscript received September 3, 2007
First revision received March 19, 2008
Final revision received April 9, 2008
Accepted on April 9, 2008

APPENDIX

Each article was dichotomously coded for reporting practices on items listed below. Each test listed had a different column. To identify reporting of inferential statistics, the test conducted was coded as either 1 = *reported inferential statistic* or 0 = *did not report inferential statistic*. If the test was not conducted, the test category was left blank.

Sample demographics (ethnicity, gender, age)
Comparison of sample to population (ethnicity, gender, age, other characteristics)
Method of participant selection
Use of national data set
Use of survey (results based on survey, survey conducted by researcher, survey available)
Factor analysis (rotation method, pattern/structure coefficients)
Reporting on reliability (data in hand, reliability of other studies, incorrectly reported reliability as the reliability of the test)
Means and standard deviations/variance
Effect size (interpreted, reported)
Confidence interval
Information needed for replication (cross-validation, bootstrap, jackknife)
Exact *p* value
Statistical significance (stated)
Figures (type of figure)

Test conducted: ANOVA, ANCOVA, *t* test, paired *t* test, correlation analysis, multiple regression, stepwise regression, HLM, OLS, multilogistic regression, commonality analysis, cross-tabs, chi-square, correlations, MANOVA, post hoc ANOVAS, MANCOVA, DDA or PDA, log-linear analysis, nonparametric test (specify), parametric test (specify)

Note. HLM = hierarchical linear modeling; OLS = ordinary least squares; DDA = descriptive discriminant analysis; PDA = predictive discriminant analysis.