Journal of Urban Mathematics Education
December 2020, Vol. 13, No. 2, pp. 42–59

GUEST EDITORIAL

Key Challenges and Some Guidance on Using Strong Quantitative Methodology in Education Research

Robin K. Henson
University of North Texas

Genéa K. Stewart
University of North Texas

Lee A. Bedford
University of North Texas

Educational researchers often struggle to draw conclusions from quantitative methods in ways that honor local contexts (Casad et al., 2017; Gutiérrez, 2002; Valero, 2008). Some in the field of mathematics education have called for more specificity when developing statistical models and assessments that can inform the field about *why, how* (Cai et al., 2019), and *for whom* (Adler et al., 2005; Connolly et al., 2018) interventions and assessments work, even down to the lesson level (Cai et al., 2020). Such concerns elevate the need to evaluate carefully, with strong methods, which curricular interventions and assessments should be brought to scale.

In educational research, idiographic realities often detract from our ability to see consistent results over time and across sites. Berliner (2002, p. 19) pointed out that a "ubiquity" of interactions is what helps to make educational research the hardest science of all. Still, the field seeks to implement research designs that help to identify practical effects of curricular interventions despite any variations between individual students, teachers, schools, and districts in which the study is conducted. Appropriate use of any methodology requires extensive, sound decision making to warrant conclusions drawn from the method.

Indeed, it can be considered somewhat of an art to parse out the influences of situational, organizational, and environmental factors in order to assess impact. Attention to good quantitative practice is worthwhile however painstaking. Strong methodological practices may help mathematics teachers and practitioners avoid lamenting the "whiplash" of starting over every few years due to frequent educational reforms (Cai et al., 2020, p. 134), which can be based on weak, over-generalized evidence.

ROBIN K. HENSON is Chair and University Distinguished Teaching Professor in the Department of Educational Psychology at the University of North Texas, 1155 Union Circle #311335 Denton, TX 76203-5017; e-mail: robin.henson@unt.edu. His research interests include applied general linear model analyses, measurement, and self-efficacy theory.

GENÉA K. STEWART is a PhD student and Teaching Fellow in the Department of Educational Psychology at the University of North Texas; email: genea.stewart@unt.edu. Her research interests include multilevel modeling, issues in educational equity, mental health, and help-seeking attitudes in college students.

LEE A. BEDFORD is a PhD candidate in the Department of Educational Psychology at the University of North Texas; email: leebedford@my.unt.edu. His research interests include measurement, trauma, and military psychology.

The thread of *context* should be woven through all stages of the research process, from design to reporting, and this focus should include whether the methods used are appropriate for the purpose of the research. Unfortunately, methodological errors (or perhaps less-than-optimal decisions) can be common. The peer-review process helps adjudicate research quality, but it can still result in research with common flaws. Though typically highly knowledgeable in their respective fields, reviewers are human. They are often very busy researchers bound by time, competing obligations, and varied methodological expertise. Moreover, the field of quantitative methodology is constantly evolving (see Aiken et al., 2008; Henson, 2006; Hughes et al., 2010; Thompson, 1999) regardless of the misconception that the field is static. All of these factors, and others, contribute to the importance of maintaining a current and appropriate understanding of quantitative methodology while considering the context surrounding its use and interpretation. Failure to do so may result in flawed research practices that can distort applications of theory, misinform policy and budgetary decisions, or even result in negative research funding decisions. As we continue to refine our understandings of best practices for quantitative methodology, it is incumbent on researchers, reviewers, and journal editors to stay well-versed in new developments.

Toward this end, the purpose of the current article is to review several common areas of focus in quantitative methods with the hope of providing *Journal of Urban Mathematics Education* (*JUME*) readers with some guidance on conducting and reporting quantitative analyses. Our intent is to challenge and stimulate strong methodological thinking. After providing some background for the needed discussion, we will review briefly the nature of recent *JUME* articles and then comment on several quantitative issues that deserve our attention while referring readers to resources for more comprehensive treatments.

Where are We Now? A Brief Review of Some Common Errors

Unfortunately, examinations of analytic and reporting practices underscore the prevalence of errors and omissions. Kesselman and colleagues (1998) conducted a comprehensive review of 17 education and behavioral science research journals for articles that contained at least one of the following: analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), and analysis of covariance (ANCOVA). They found that statistical assumptions are often not reported or are even violated, effect sizes are rarely reported, and sample sizes are not regularly based on power analyses (Kesselman et al., 1998). Both inadequate sample sizes and non-random sampling introduce bias in the interpretation of results in the form of increased sampling error, which decreases the accuracy of findings (Tabachnick & Fidell, 1996). In a broader review focused specifically on education research, Zientek et al. (2008) examined 174 articles cited by the American Educational Research

Association Panel on Research and Teacher Education (see Cochran-Smith & Zeichner, 2005), finding only 13% of the articles reported score reliability, 4% reported confidence intervals, and 39% reported effect sizes. Furthermore, in a review of the *Journal of Applied Psychology,* Courville and Thompson (2001) found that 94% of articles contained discrepancies between beta weights and structure coefficients when ranking predictors in regression analyses. The authors also highlighted other common errors and misinterpretations of regression analyses.

There are common mistakes to be found in studies that use exploratory factor analyses (EFAs) as well. Henson and Roberts (2006) conducted a review of 60 articles that used EFAs and found that some studies contained less than the recommended sample size (median $N = 267$); Tabachnick and Fidell (1996) would consider this to be below the minimum sample size for an EFA. Many studies also contained less than the recommended amount of variance explained by the factors, and often researchers likely did not extract the correct number of factors. Furthermore, 65% of studies did not report which matrix of association was analyzed, 13% did not report the extraction method used, and nearly 57% used the default extraction method in their EFAs.

Henson et al. (2010) identified and emphasized deficiencies in quantitative and research methods training in education doctoral programs that may lead to usage errors and reporting problems in research articles. They also argued that researchers rely too heavily on traditional research designs and statistical analyses, resulting in limited learning and application of new advances in quantitative methodology. The authors suggested several ideas for advancement, such as additional training and consulting with methodologists early in the process when designing studies.

The current article can be reasonably considered as one small educational step in that direction. Our approach is holistic in the sense that we fully comment on the research process in multiple areas (e.g., design, data analysis, and reporting). It is impossible to provide comprehensive guidance in one article, but we address some key challenges and best practices in quantitative research in the following domains: causal inferences, measurement, handling missing data, testing for assumptions, addressing nested data, and evidence for outcomes. These domains were selected based on discrepancies consistently identified in methodological reviews of educational research (Aiken et al., 2008; Connolly et al., 2018; Courville & Thompson, 2001; Enders, 2010; Henson et al., 2010; Henson & Roberts, 2006; Kesselman et al., 1998; Peugh & Enders, 2004; Vacha-Haase et al., 1999; Zientek et al., 2008).

An Empirical Review of Recent *JUME* Methods

To provide a baseline for the discussion, we reviewed all *JUME* publications categorized as research articles ($N = 24$) from 2014–2017 to evaluate the types of research methods typically employed in the journal. The 2018 volume was not

included because it consisted of reprints from 2008 as a 10-year follow-up. The 2019 volume was not yet available online. All articles classified under the editorial, commentary, response commentary, or research impact categories were excluded.

Regular readers of *JUME* would likely not be surprised that the journal has historically tended toward publication of qualitative-oriented articles. Fully 75% of the articles reviewed could be classified as qualitative, whereas only 16.7% utilized quantitative designs and 8.3% used some form of a mixed methods approach. Table 1 provides a summary of the review and also illustrates the types of analyses employed. Because our focus is on quantitative research, we did not delineate types of qualitative approaches.

Table 1

Methodological Approach in *JUME* Research Articles (2014–2017)

Type of Method	*n*	Primary Method
Qualitative methodology	18	✓
Quantitative and mixed methodology	6	
Cross-tabulations	1	
Paired sample t-tests	1	✓
Independent samples t-tests	1	
Propensity score matching	1	✓
Multiple regression	1	✓
Meta-analysis	1	✓
Item response theory	1	
Analytic weights	2	
Chi-square	1	
Coefficient alpha	3	
Descriptive statistics	5	

Although the nature of *JUME*'s historical focus may have gravitated toward a qualitative emphasis, the policies and procedures of the journal indicate an interest in publishing "data based qualitative and quantitative studies, action research, research syntheses, integrative reviews and interpretations of research literature" (Journal of Urban Mathematics Education, n.d.-a). In many cases, qualitative methods allow for strong idiographic exploration of phenomena at the local level (Berliner, 2002; Demerath, 2006). However, appropriate integration of strong quantitative methodology could broaden the journal's goals of publishing data-based research and promoting diverse applications to "foster a transformative global academic space in mathematics" (Journal of Urban Mathematics Education, n.d.-b). Of course, in any study, the research question should drive the methods used. If appropriately applied and done well, quantitative methods can help bolster research inquiry with measurable and distinct evidence for effects of interventions and other correlational questions.

Some Key Challenges and Comment on Good Practice

Below are suggestions and resources for addressing problems in the following six quantitative domains: causal inferences, measurement, handling missing data, testing for assumptions, addressing nested data, and evidence for outcomes. These suggestions are based on previous methodological reviews in the fields of education (e.g., Henson et al., 2010) and psychology (e.g., Aiken et al., 2008). It is our hope that researchers in the field of urban mathematics education will heed warnings identified in these related fields in order to produce higher quality manuscripts. Recent empirical studies in urban mathematics education are referenced throughout this paper to provide support for methodological recommendations.

Causal Inferences

Researchers are often concerned with making causal claims between variables. The field of mathematics education, in particular, would benefit from an increased understanding of causal processes within the classroom that influence instructional approaches across explicit conditions (Cai et al., 2019; Maxwell, 2004), such as student perceptions of equity and access to participation (Vogler et al., 2018).

It is not uncommon to find statements of causal inference and generalization in qualitative studies. Although the claims may be presented tentatively, causal inferences are either explicit or implied far more often than is warranted by our designs — whether qualitative or quantitative. Researchers may take for granted the following essential elements needed for causation: (a) the independent variable and dependent variable must be correlated, (b) the independent variable should take temporal precedence, or come first, over the dependent variable, and (c) there must be no effect of extraneous variables (Shadish et al., 2002). Therefore, the ideal approach for

researchers to determine causation is through manipulation of variables where any outcome differences at post are attributable to the treatment and *not* pre-existing systematic differences between groups (e.g., self-selection). Experimental designs are considered very strong because they allow the researcher to observe "phenomena which are made to occur in strictly controlled situations in which one or more variables are varied and the others are kept constant" (Zimney, 1961, p. 18). Random assignment of participants to conditions helps researchers rule out rival or competing hypotheses because theoretically confounding variables are leveled through randomization. It follows that experimental research lends itself well to formation of causality arguments. Still, experimental research is not often feasible for much of educational research, or even desirable.

Quasi-experimental studies are not as strong as experimental but are still preferred over correlational studies because conditions can be manipulated so that (a) covariance between the intervention and outcome can be observed and (b) the intervention precedes the effect (Johnson & Christensen, 2019). However, quasi-experimental studies fall short of the third causal criteria, as they do not fully account for confounding variables. This is often because researchers cannot randomly assign students to conditions (e.g., classrooms or schools) of their choosing or have no ability to include a control group. Collecting data necessary to address competing explanations can strengthen causal arguments with this design.

In educational research, randomized control trials (RCTs), studies in which individuals or groups of individuals are randomly assigned to treatment conditions, are becoming an increasingly popular approach toward the development of evidence-based practices and theories of change (Connolly et al., 2018). However, to avoid overgeneralization of cause and effect claims, Connolly et al. (2018) advised that RCTs in education should involve some sub-group analyses and incorporate process evaluations as a component of the research.

Finally, in observational studies where random assignment is simply not possible due to either program criteria or the practical logistics of the setting, propensity score matching (PSM) offers a valuable alternative (Austin, 2011; Henson et al., 2010; Morgan et al., 2010). PSM seeks to reduce selection bias by approximating a randomized experiment with "treatment" and "control" groups based on participant covariates and evaluating whether differences are likely due to treatment (Austin, 2011; Rosenbaum & Rubin, 1983). Thus, the PSM process can serve as an analog to the random experiment and is superior to drawing conclusions from observation alone. One of the major drawbacks to adopting PSM is that it fails to balance unobserved, or unmeasured, characteristics in the statistical model and it rests on the core assumption that all confounders are measured (Austin, 2008; Hill, 2008). RCTs work better to balance measured and unmeasured covariates across intervention groups (Austin, 2008). Regardless of method, researchers should take care to qualify their findings with clear indications as to whether any causal claims are supported. As an

example, Howard et al. (2015) demonstrated good practice in tempering causal language (i.e., qualifying conclusions) within a PSM study using a secondary longitudinal dataset:

> It is important to note that it cannot be inferred that the more negative psychological scores were necessarily caused by the failing grades. It is feasible that the students failed because they already had more negative dispositions toward mathematics, or conversely that negative dispositions emerged following the failure. Regardless of the direction of causality (if any), the existence of more negative dispositions towards mathematics indicates that psychological dispositions are somehow associated with mathematics performance in the eighth grade, and placing these students in advanced mathematics courses without addressing these dispositions may be inadequate in terms of the support they may need. (p. 54)

The reporting in this particular *JUME* manuscript (included in our empirical review) was comprehensive and informed. Further elaboration on limitations of design in papers such as this will help to move the mathematics education field toward an even deeper understanding of the extent to which we can draw causal inferences across study designs.

Measurement Issues

Researchers need to either use instruments that yield scores with strong psychometric properties or create measures themselves that yield scores with strong reliability and validity. Further, researchers should use caution when creating their own measures (e.g., surveys). Scores from every measure should be deemed valid and reliable prior to being used in published research, as poorly written surveys that may appear to only have face validity can weaken the foundation of entire bodies of research (Borsboom et al., 2004; Lissitz & Samuelson, 2007). Measurement is often the Achilles' heel of quantitative research. If we do not measure what we are interested in well, then it may not matter what else we do in the study!

In a classical test theory framework, test scores are considered reliable when the test's "observed scores are highly correlated with its true scores" (Allen & Yen, 1979, p. 72). This theoretical concept is often operationalized and assessed with test-retest reliability correlations and, most commonly, with coefficients measuring internal consistency (i.e., correlation) between items (Hogan et al., 2000). Henson (2001) provides a primer on the meaning and interpretation of internal consistency reliability coefficients, such as coefficient alpha (see Cronbach, 1951).

Unfortunately, researchers often ignore reliability or incorrectly assume that a measure will yield reliable scores in a current study just because it has in prior samples. Vacha-Haase et al. (1999) reviewed 839 articles in three counseling and psychology journals and found that only 35.6% of the articles reported reliability coefficients for the data being analyzed. These coefficients are essential for evaluating good

measurement because scores, not tests themselves, are either reliable or unreliable (Vacha-Haase et al., 2002).

Regarding measurement validity (i.e., measuring what you claim to be measuring), researchers should expect evidence for construct, content, convergent, and discriminant validity before assuming that scores on a measure may be valid in a current or future sample (Borsboom et al., 2004; Cronbach & Meehl, 1955; Lissitz & Samuelson, 2007). In order to consider a measure good enough to be included in one's research study, there should be sufficient evidence for score validity from prior work in similar samples. Failure to evaluate and report the validity evidence for scores on a measure amounts to rolling the dice on whether one's obtained scores in a current study will mean anything at all. Score reliability and validity cannot be assumed even if an instrument was previously published in a reputable journal and used widely in the literature (see Henson et al., 2001).

Not only do researchers need to select measures that will yield reliable and valid scores, they must also be aware of measurement invariance between samples or groups (i.e., there may be systematic biases in measures based on the sample being tested; Millsap, 2011). Researchers often believe that "reliability coefficients from previous samples or test manuals are psychometrically applicable for their current published work," but this is simply not true in an absolute sense (Vacha-Haase et al., 2002, pp. 563–565). Factor structures of the measures should be assessed (Henson & Roberts, 2006), and researchers should "investigate the invariance of the measures implemented before comparing the results from the measure in a study" (Henson et al., 2010, p. 234). Millsap (2011) provides a review of measurement invariance testing.

Of course, understanding these measurement implications can be difficult when not regularly part of researcher training. In a survey of psychology doctoral programs in the United States and Canada, Aiken et al. (2008) found that only 64% of all departments provided a doctoral course in measurement. Furthermore, the researchers found that 40% of departments taught sections that included item response theory (IRT) as an element every 2 years, with only 9% teaching a full semester of IRT. IRT is an advanced measurement approach that allows for item parameters and test-taker ability to be assessed on the same scale (Reise et al., 2005). For example, one study focused on understanding the Flynn effect was able to separate out general intelligence, an ability parameter, from item parameters (e.g., discrimination and difficulty) on a mathematics achievement test (Beaujean & Osterlind, 2008). IRT is a key method in modern test development (Embretson & Reise, 2000), and recent studies have utilized IRT to create and validate scores on mathematics assessments (e.g., the Probabilistic Reasoning Scale; see Primi et al., 2017; the Abbreviation Math Anxiety Scale; see Sadiković et al., 2018).

Missing Data Techniques

Missing data is a common occurrence in educational research, yet modern missing data imputation techniques, such as multiple imputation (MI) and full information maximum likelihood estimation (FIML; Schafer & Graham, 2002), are not regularly taught in doctoral programs (Aiken et al., 2008). Researchers often address missing data by either removing the cases that contain missing values (i.e., listwise deletion) or with mean, median, or regression imputation of the missing values (Peugh & Enders, 2004). Listwise deletion leads to biased parameter estimates if the data are not missing completely at random (MCAR; Little, 1988; Peugh & Enders, 2004). Furthermore, even if the data are MCAR, mean imputation can produce biased parameter estimates (e.g., inaccurate R^2 effect sizes in regression models). Morales-Chicas and Agger's (2017) math achievement study highlighted the limitations introduced when data are missing and reported the listwise deletion of cases. A description of the missing data and a statement about whether the data were MCAR would strengthen the reporting and allow readers to better interpret the final results.

Newer methods of missing data imputation, such as MI and FIML, can produce unbiased estimates for missing values that are either MCAR or missing at random. It should be noted that values that are missing not at random (i.e., systematically missing) will tend to produce biased parameter estimates regardless of which missing data technique is utilized (Peugh & Enders, 2004). Howard et al. (2015) provided transparent and detailed reporting of their missing data decision-making process wherein the authors compared results from multiple imputation data and the original data but ultimately used the original data due to negligible differences between the two:

> Little's MCAR statistic (SPSS Missing Values 22.0) revealed that the missing data met the assumption of MCAR, $\chi^2(39) = 52.84$, $p = .07$. There were no systematic patterns of missing data when compared to the observed values for all of the matched covariates, the prior mathematics assessment and psychological measure scores, and the college-bound variables. (p. 48)

FIML uses a maximum likelihood estimation instead of the least squares estimation that other imputation techniques utilize. MI creates multiple imputed datasets and pools the estimates in a two-step process (Enders, 2010). The primary benefit of both techniques is that they require less strict assumptions regarding the missing values (Peugh & Enders, 2004). FIML is slightly more accurate (Schafer & Graham, 2002), but MI can be more versatile in the imputation process (Peugh & Enders, 2004). See Enders (2010) for a full guide to missing data techniques.

Testing for Assumptions

Because inferential statistical analyses use data from samples to generalize to populations, each analysis is accompanied by a set of assumptions (Cohen et al.,

2003; Tabachnick & Fidell, 1996). Thus, the key assumptions need to be tested *and* reported in published articles. In the event that an analysis' assumptions are violated, researchers should evaluate the implications based on the severity of the violation and consider either data transformations, seek out correction methods, or switch to nonparametric tests to ensure the most accurate results (Osborne, 2013).

Researchers should also be aware that assumptions change based on the analysis used. For instance, the assumption of multivariate normality is a prerequisite for an analysis that involves the simultaneous prediction of multiple outcomes (Henson, 1999). In turn, homogeneity of variance is assumed for between-group comparisons, and sphericity is assumed when testing change within subjects over time (Tabachnick & Fidell, 1996). The assumption of homoscedasticity, wherein all values of X share the same scatter around a regression line, should be met in order to draw accurate conclusions from analyses under the general linear model, such as multiple regression (Cohen et al., 2003). Yet, in a review of 61 articles that utilized between-subject analyses, Kesselman et al. (1998) found that only one study reported testing for both normality and homogeneity of variance. Furthermore, Onwuegbuzie and Daniel (2005) conducted a review of articles submitted to *Research in the Schools* and noted that 91% of submitted manuscripts did not discuss model assumptions. It is not uncommon for reporting of these assumptions to be omitted in published multiple regression articles in the field of mathematics education (e.g., Irvin et al., 2017; Lee, 2018; Morales-Chicas & Agger, 2017; Smith & Hoy, 2007). Reporting assumptions is essential to good quantitative practice, for when we report results without checking or meeting the assumptions, we risk publishing results that are not replicable (Kesselman et al., 1998).

Addressing Nested Data

The complex research questions that drive our analyses of program effectiveness are often concerned with data at multiple levels (e.g., classrooms, schools, districts, and states). People or students in these naturally occurring hierarchical groupings, also referred to as clusters, often share variance on outcomes because of their common experience, setting, and so forth. Therefore, when researchers treat clustered data as independent, they are at an increased risk of making Type I errors (Ferron et al., 2008). Instead, we can consider the data as *nested* (i.e., students at level 1 nested within schools at level 2) and adjust our model to account for these cluster effects (e.g., non-independence of data). These multilevel models, also referred to as hierarchical linear models and variously as mixed models, mixed effect models, and random coefficient models (McCoach, 2010; Raudenbush & Bryk, 2002), allow for interpretation of both *fixed effects*, which may be estimates of slopes and intercepts that do not vary by cluster, and *random effects*, which are allowed to vary by organization unit or by individual (Woltman et al., 2012).

Multilevel research questions related to cross-sectional analyses often seek to account for variability stemming from student-level differences and any cluster effects of the aggregate unit (e.g., Does class size moderate the relationship of student self-esteem and mathematics achievement?). To illustrate the benefits of multilevel modeling for addressing complex research questions in mathematics education, consider a study by Young (1997). The author employed multilevel modeling to decompose science and mathematics achievement by simultaneously investigating predictors at the teacher/school level (e.g., student support or mission consensus), classroom level (e.g., student cohesiveness or task orientation), and student level (e.g., self-concept and satisfaction). In that study, class and school effects were present, but ultimately student self-concept explained the most variance in achievement. This important finding informs educators' ongoing engagement with students. In urban mathematics education, Lekwa et al. (2019) also used multilevel modeling, this time with students nested in classrooms, to measure "the predictive relationship between teacher practices, as measured by the CSAS-O, and gains in student achievement in reading and mathematics" (p. 15).

Use of multilevel modeling in longitudinal designs can additionally allow exploration of differences in student growth rates (e.g., To what extent do boys and girls differ in their rate of change in self-confidence based on program involvement?). Multilevel modeling also offers the advantage of increased statistical power to detect growth effects when student data is missing at various waves (Kwok et al., 2008).

In order to determine the degree of dependence of the data and provide some preliminary guidance on whether a multi-level model would even be appropriate, researchers can compute an intraclass correlation coefficient (ICC). The ICC provides a proportion of variance that can be explained by hierarchical groups, and it is computed as a ratio of between group variance divided by the sum of the between group variance and within group variance in a model with no specified predictors (Raudenbush & Bryk, 2002, p. 71). The ICC can also be more simply explained as the degree of similarity we might expect for any two randomly selected student scores within the same organizational unit. In an instructive chapter on good reporting practices in hierarchical linear modeling, McCoach (2010) explained, "An ICC of 0 indicates independence of observations, and any ICC above 0 indicates some degree of dependence in the data" (p. 134). For instance, Matthews' (2018) study on urban adolescents' cognitive flexibility and views of mathematics exhibits use of the ICC to make statistical decisions:

> The intraclass correlation coefficient for year-end attainment value was .042, which indicated that the large majority of variation in attainment value existed between students and very little, 4.2%, between classrooms. Little variation at the classroom level reduces the need for multilevel modeling (Bryk & Raudenbush, 1992). (p. 6)

Evidence for Outcomes

Perhaps the most central question regarding quantitative analyses is the degree to which they provide information about evidence for study outcomes. Any study can typically yield multiple pieces of information that can be consulted as evidence that helps communicate the story found in the data. It is generally important to consider several elements when evaluating outcomes, such as effect size (i.e., practical significance), confidence intervals, power, and statistical significance.

Regarding traditional null hypothesis statistical significance testing, Onwuegbuzie and Daniel (2005) found that 86% of studies did not include any discussion of a priori or post-hoc statistical power analyses, while 14% clearly lacked adequate statistical power. If null hypothesis statistical significance testing is going to be used, sufficient power is necessary to provide context for interpreting obtained *p* values (Cohen, 1983). Furthermore, 33% of the studies had omitted one or more appropriate statistics (e.g., degrees of freedom, *p* value), and 33% confused statistical with practical significance. The majority (65%) of articles had no discussion of limitations and legitimacy of findings.

These omissions are not uncommon in the literature, and they reflect a general lack of understanding of the importance of providing clear evidence to support claims regarding study outcomes. For example, it is now commonly accepted that multiple pieces of evidence should be reported and that this should include effect size interpretation and confidence intervals beyond traditional statistical significance testing. Henson (2006) argued for stronger meta-analytic thinking across the literature when evaluating study outcomes with emphasis on reporting and interpreting effect sizes. Young et al. (2019) demonstrated careful interpretation of results in their study on effects of urban mathematics teachers' professional development and stated, "although isolated effect size results suggest an overall positive outcome for the professional development, meta-analytic thinking can contextualize the results and provide a broader interpretation of the professional development effectiveness" (p. 322). Along with other meta-analytic considerations (see Cumming & Finch, 2005; Quintana & Minami, 2006), Henson (2006) suggested reporting confidence intervals around obtained effects (see also Thompson, 2002).

Indeed, the most recent *Publication Manual of the American Psychological Association* noted that "for readers to appreciate the magnitude or importance of a study's findings, it is recommended to include some measure of effect size in the Results section" (American Psychological Association [APA], 2020, p. 89). Past language also stressed, "it is almost always necessary to include some measure of effect size in the Results section" (APA, 2010, p. 34). Effect sizes can be reported in original units (e.g., a regression slope) or standardized units (e.g., Cohen's *d* value), and researchers should use their best judgement about which approach is preferred for better interpretability when communicating magnitude of effect (APA, 2020).

Common effect sizes include variance-accounted-for measures (e.g., R^2, η^2) and standardized mean differences (e.g., Cohen's *d*, Hedges' *g*).

Finally, in many analyses in which interesting effects are observed with sufficient evidence for the study outcome, researchers must evaluate the role of variables in the model. For example, an interesting multiple regression model would necessitate interpretation of which predictors were most impactful in explaining variability in the outcome. In such cases, a traditional approach would focus only on the standardized coefficients, or perhaps unstandardized versions, of the predictors. However, comprehensive researchers will also include interpretation of structure coefficients in conjunction with beta weights to evaluate potential impacts of multicollinearity, the relative strength of predictors, and possible suppressor effects. Readers are referred to Courville and Thompson (2001), Henson (2002), and Kraha et al. (2012) for detailed explanations of structure coefficients.

Summary

Our primary goal as researchers is to produce good research that implements strong research designs with results that are understood by both researcher and practitioner in a way that can be practically applied. Although not always the case, we also often care about whether outcomes are generalizable and replicable. Before particular methods are considered and employed for these ends, however, we should be asking strong research questions that are grounded in theory and contextualized in appropriate settings. Only then should our consideration turn to the appropriate use of methodology to help answer our questions.

This article highlighted key challenges and some best practices we find in quantitative research in several areas. Decisions made in these domains should be reported in manuscripts whenever possible for transparency. We also offered resources to pursue in order to strengthen quantitative research designs and practices. Editors and peer-reviewers are essential to this process, as they are gatekeepers to the publication of good research. When relevant, peer-reviewers should understand these quantitative practices and apply that understanding with thoughtful and detailed feedback to authors. It is also very helpful to provide resources and relevant citations to assist authors with the publication process.

Aiken et al. (2008) noted glaring weaknesses in the quantitative training provided in psychological doctoral programs. Many programs do not teach modern measurement analyses, advanced techniques for handling missing data, or complex inferential analyses. Henson and Williams (2006) found similar outcomes in education doctoral programs. Conducting quality quantitative research is obviously a complex process, but training in terminal degree programs is certainly a piece of the puzzle. It is also reasonable to think that the professional development of researchers would play a role in whether or not quality methodology is used in the literature. Like

other fields, quantitative practice changes and improves through methodological and technological advancements. If we are going to conduct and publish quantitative-oriented research, we need to do it well, and doing it well requires that we make good, contextualized decisions throughout the entire research process.

References

Adler, J., Ball, D., Krainer, K., Lin, F., & Novotna, J. (2005). Reflections on an emerging field: Researching mathematics teacher education. *Educational Studies in Mathematics, 60*(3), 359–381. https://doi.org/10.1007/s10649-005-5072-6

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*(1), 32–50. https://doi.org/10.1037/0003-066X.63.1.32

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole Publishing Company.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.).

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049. https://doi.org/10.1002/sim.3150

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 Children and Young Adults data. *Intelligence*, *36*(5), 455–463. https://doi.org/10.1016/j.intell.2007.10.004

Berliner, D. C. (2002). Comment: Educational research: The hardest science of all. *Educational Researcher*, *31*(8), 18–20. https://doi.org/10.3102/0013189X031008018

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. http://doi.org/10.1037/0033-295X.111.4.1061

Cai, J., Morris, A., Hohensee, C., Hwang, S., Robison, V., Cirillo, M., Kramer, J., & Hiebert, J. (2019). Posing significant research questions. *Journal for Research in Mathematics Education*, *50*(2), 114–120. http://doi.org/10.5951/jresematheduc.50.2.0114

Cai, J., Morris, A., Hohensee, C., Hwang, S., Robison, V., Cirillo, M., Kramer, S. L., Hiebert, J., & Bakker, A. (2020). Addressing the problem of always starting over: Identifying, valuing, and sharing professional knowledge for teaching. *Journal for Research in Mathematics Education*, *51*(2), 130–139. https://doi.org/10.5951/jresematheduc-2020-0015

Casad, B. J., Hale, P., & Wachs, F. L. (2017). Stereotype threat among girls: Differences by gender identity and math education context. *Psychology of Women Quarterly*, *41*(4), 513–529. https://doi.org/10.1177%2F0361684317711412

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

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum.

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, *60*(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement, 61*(2), 229–248. https://doi.org/10.1177/0013164401612006

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. http://doi.org/10.1037/h0040957

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*(2), 170–180. http://doi.org/10.1037/0003-066X.60.2.170

Demerath, P. (2006). The science of context: Modes of response for qualitative researchers in education. *International Journal of Qualitative Studies in Education*, *19*(1), 97–113. https://doi.org/10.1080/09518390500450201

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data*. Information Age Publishing.

Gutiérrez, R. (2002). Enabling the practice of mathematics teachers in context: Toward a new equity research agenda. *Mathematical Thinking and Learning*, *4*(2–3), 145–187. https://doi.org/10.1207/S15327833MTL04023_4

Henson, R. K. (1999). Multivariate normality: What is it and how is it assessed? *Advances in Social Science Methodology*, *5*, 193–211.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*(3), 177–189. https://doi.org/10.1080/07481756.2002.12069034

Henson, R. K. (2002, April 1–5). *The logic and interpretation of structure coefficients in multivariate general linear model analyses* [Paper presentation]. Annual Meeting of the American Educational Research Association, New Orleans, LA, United States.

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, *34*(5), 601–629. https://doi.org/10.1177/0011000005283558

Henson, R. K., Hull, D. M., & Williams, C. S. (2010). Methodology in our education research culture: Toward a stronger collective quantitative proficiency. *Educational Researcher, 39*(3), 229–240. https://doi.org/10.3102/0013189X10365102

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61*(3), 404–420. https://doi.org/10.1177/00131640121971284

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*(3), 393–416. https://doi.org/10.1177/0013164405282485

Henson, R. K., & Williams, C. (2006, April 7–11). *Doctoral training in research methodology: A national survey of education and related disciplines* [Paper presentation]. Annual Meeting of the American Educational Research Association, San Francisco, CA, United States.

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*. *Statistics in Medicine*, *27*(12), 2055–2061. https://doi.org/10.1002/sim.3245

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*(4), 523–531. https://doi.org/10.1177/00131640021970691

Howard, K. E., Romero, M., Scott, A., & Saddler, D. (2015). Success after failure: Academic effects and psychological implications of early universal algebra policies. *Journal of Urban Mathematics Education*, *8*(1). https://doi.org/10.21423/jume-v8i1a248

Hughes, G. D., Onwuegbuzie, A. J., Daniel, L. G., & Slate, J. R. (2010). APA Publication Manual changes: Impacts on research reporting in the social sciences. *Research in the Schools*, *17*(1), viii–xix.

Irvin, M., Byun, S. Y., Smiley, W. S., & Hutchins, B. C. (2017). Relation of opportunity to learn advanced math to the educational attainment of rural youth. *American Journal of Education*, *123*(3), 475–510. https://doi.org/10.1086/691231

Johnson, R. B., & Christensen, L. (2019). *Educational research: Quantitative, qualitative, and mixed approaches*. SAGE.

Journal of Urban Mathematics Education. (n.d.-a). *Policies and procedures*. Retrieved November 1, 2019, from https://jume-ojs-tamu.tdl.org/jume/index.php/jume/policiesandprocedures

Journal of Urban Mathematics Education. (n.d.-b). *About the journal*. Retrieved November 1, 2019, from https://journals.tdl.org/jume/index.php/jume/about

Kesselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*(3), 350–386. https://doi.org/10.3102/00346543068003350

Kraha, A., Turner, H., Nimon, K., Zientek, L., & Henson, R. (2012). Tools to support interpreting multiple regression in the face of multicollinearity. *Frontiers in Psychology*, *3*, 44. https://doi.org/10.3389/fpsyg.2012.00044

Kwok, O., Underhill, A., Berry, J. W., Luo, W., Elliott, T., & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures. *Rehabilitation Psychology, 53*(3), 370–386. https://doi.org/10.1037/a0012765

Lee, L. S. (2018). Success of online mathematics courses at the community college level. *Journal of Mathematics Education*, *11*(3), 69–89. https://doi.org/10.26711/007577152790033

Lekwa, A. J., Reddy, L. A., Dudek, C. M., & Hua, A. N. (2019). Assessment of teaching to predict gains in student achievement in urban schools. *School Psychology*, *34*(3), 271–280. https://doi.org/10.1037/spq0000293

Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437–448. https://doi.org/10.3102/0013189X07311286

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Matthews, J. S. (2018). When am I ever going to use this in the real world? Cognitive flexibility and urban adolescents' negotiation of the value of mathematics. *Journal of Educational Psychology*, *110*(5), 726–746. http://doi.org/10.1037/edu0000242

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, *33*(2), 3–11. https://doi.org/10.3102%2F0013189X033002003

McCoach, D. B. (2010). Hierarchical linear modeling. In G. R. Hancock, R. O. Mueller, & L. M. Stapleton (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 123–140). Routledge.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Morales-Chicas, J., & Agger, C. (2017). The effects of teacher collective responsibility on the mathematics achievement of students who repeat algebra. *Journal of Urban Mathematics Education*, *10*(1), 52–73. https://doi.org/10.21423/jume-v10i1a287

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *Journal of Special Education*, *43*(4), 236–254. https://doi.org/10.1177/0022466908323007

Onwuegbuzie, A. J., & Daniel, L. G. (2005). Evidence-based guidelines for publishing articles in *Research in the Schools* and beyond. *Research in the Schools*, *12*(2), 1–11.

Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data.* SAGE.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556. https://doi.org/10.3102/00346543074004525

Primi, C., Morsanyi, K., Donati, M. A., Galli, S., & Chiesi, F. (2017). Measuring probabilistic reasoning: The construction of a new scale applying item response theory. *Journal of Behavioral Decision Making*, *30*(4), 933–950. https://doi.org/10.1002/bdm.2011

Quintana, S. M., & Minami, T. (2006). Guidelines for meta-analyses of counseling psychology research. *The Counseling Psychologist, 34*(6), 839–877. https://doi.org/10.1177/0011000006286991

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*(2), 95–101. https://doi.org/10.1111/j.0963-7214.2005.00342.x

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Sadiković, S., Milovanović, I., & Oljača, M. (2018). Another psychometric proof of the Abbreviated Math Anxiety Scale usefulness: IRT analysis. *Primenjena Psihologija*, *11*(3), 301–323. https://doi.org/10.19090/pp.2018.3.301-323

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Smith, P. A., & Hoy, W. K. (2007). Academic optimism and student achievement in urban elementary schools. *Journal of Educational Administration*, *45*(5), 556–568. https://doi.org/10.1108/09578230710778196

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). Pearson.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*, *9*(2), 165–181. https://doi.org/10.1177/095935439992006

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25–32. https://doi.org/10.3102/0013189X031003025

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*(4), 562–569. https://doi.org/10.1177/0013164402062004002

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education*, *67*(4), 335–341. https://doi.org/10.1080/00220979909598487

Vogler, A. M., Prediger, S., Quasthoff, U., & Heller, V. (2018). Students' and teachers' focus of attention in classroom interaction — Subtle sources for the reproduction of social disparities. *Mathematics Education Research Journal*, *30*(3), 299–323. https://doi.org/10.1007/s13394-017-0234-2

Valero P. (2008). In between the global and the local: The politics of mathematics education reform in a globalized society. In B. Atweh, A. C. Barton, M. C. Borba, N. Gough, C. Keitel, C. Vistro-Yu, & R. Vithal (Eds.), *Internationalisation and Globalisation in Mathematics and Science Education* (pp. 421–439). Springer. https://doi.org/10.1007/978-1-4020-5908-7_23

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52–69. https://doi.org/10.20982/tqmp.08.1.p052

Young, D. J. (1997, March 24–28). *A Multilevel Analysis of Science and Mathematics Achievement* [Paper presentation]. Annual Meeting of the American Educational Research Association, Chicago, IL, United States.

Young, J. R., Young, J., Hamilton, C., & Pratt, S. (2019). Evaluating the effects of professional development on urban mathematics teachers TPACK using confidence intervals. *REDIMAT – Journal of Research in Mathematics Education*, 8(3), 312–338. http://doi.org/10.17583/redimat.2019.3065

Zientek, L. R., Capraro, M. M., & Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: One look at the evidence cited in the AERA panel report. *Educational Researcher*, *37*(4), 208–216. https://doi.org/10.3102/0013189X08319762

Zimney, G. H. (1961). *Method in experimental psychology*. Ronald Press.
