1

# Quantitative Reporting Practices in Middle-Grades Research Journals: Lessons to Learn

**Robert M. Capraro**
*Texas A&M University*
*College Station, Texas*

**Mary Margaret Capraro**
*Texas A&M University*
*College Station, Texas*

***ABSTRACT***

*This study examines two journals specific to the middles grades where original quantitative empirical articles are published,* ***Research in Middle Level Education*** *and* ***Middle Grades Research Journal*** *to determine what quantitative statistics are used, how they are used, and what study designs are used. Important for those who write for the middle-grades community is to make a cognizant effort to adhere to recommendations that will improve the quality of what we know and how we can best situate our research findings to provide maximum information for the reader.*

## INTRODUCTION

There are many research paradigms and traditions that comprise middle grades research, this report does not in any way attempt to diminish or exalt any particular one. Several professional organizations have developed standards for quantitative reporting practices, and as of yet, there has been limited interest in attempting to establish reporting practices for the numerous rich traditions that comprise qualitative research.

It is important to examine and contextualize the strengths and weaknesses in quantitative reporting for middle-grades researchers. One important reason is that advocacy for the middle grades is often upheld by stakeholders (i.e., teachers, principals, parents, and the organizations that support these groups) and not researchers. If researchers are to make a lasting impact, likely to influence those who primarily advocate for the middle grades, they must provide authentic empirical research that is interpretable by those who will apply it. Therefore, the research published for this group must be written with attention to delivering the message to a two-tiered audience. This article carefully examines the journals specific to the middles grades where original quantitative empirical articles are published - *Research in Middle Level Education* and *Middle Grades Research Journal* - to examine what quantitative statistics are used, how they are used, and what study designs are used.

### *Rationale*

Educational research yields its greatest potential contribution to middle-grades practices when recommended reporting practices are followed (American Educational Research Association [AERA], 2006; Natriello, 2006; Wilkinson & American Psychological Association [APA] Task Force on Statistical Inference, 1999). AERA adopted *Standards* for assisting the research community in producing high quality empirical research reports. These *Standards* provide guidance to editors, reviewers, and educational researchers alike, in vetting manuscripts submitted for publication (AERA).

### *Why Examine Reporting Practices*

The middle-grades community relies on empirical reports of research to deal with recurring questions such as which grade configuration (5-8, 6-8, or 7-9, possibly others) are more likely to yield improved achievement or are self-contained classes taught by generalists preferable to rotating among teachers who are content area specialists? These are in addition to issues of emerging research topics such as social, affect and bonding, identity, technology, and cognition. To interpret results from these studies, reporting practices are important for communicating findings that will be used to influence middle-level education. In middle-grades research, we expect our findings will precipitate educational innovation and side-step known issues with the intellectual and emotional development of middle graders. While there is a need to produce middle-grades research for consumption by researchers, it is important to carefully report the research accurately and with attention to the fact that some will have insufficient background knowledge to know if there are inherent flaws in either the design or analytic aspects (Stipek, 2005) – that is, in the reporting of the research it is important to build the skills of the reader who has to interpret it. Therefore, it is essential to educate those who make important middle-grades' decisions within schools and districts (Wolk, 2007). To attain this goal, published research must include both the necessary and sufficient information for readers to evaluate claims.

## *The Lens for Examining Research*

Reporting standards facilitate research quality by articulating a set of expectations by which informed editors, their boards and reviewers, and manuscript authors can have agreed upon dialog about the appropriateness of any particular manuscript for publication. In addition to having an unbiased set of criteria by which to judge a manuscript, it is important to provide a clear understanding of how scientific claims are warranted. While the APA and the AERA have developed a generic set of standards in general for social science researchers to follow, the American Statistical Association, and the American Mathematics Association have developed sets of specific standards for their disciplines. Therefore, it is incumbent upon each discipline to develop and engender its own set of standards to ensure quality (Feuer, Towne, & Shavelson, 2002) and middle-grades research should not be left out.

There are important considerations in attaining the goal of providing useful and interpretable middle-grades research. First and foremost, our research questions should be clear. From these questions one can decide upon the design (e.g., experimental, quasi-experimental, or convenience). Then, from this design determine the research paradigm (Raudenbush, 2005). The analyses should be thoughtful and appropriately applied. It is important not to reduce primarily humanistic research endeavors of middle-grades research into a set of preprogrammed algorithms in an attempt to claim scientific rigor. All social science research, whether for middle grades or not, is about the iterative process of contemplation, reflection, and then employment of good judgment (Thompson, 2006) in making design and analytic choices. In the careful consideration of the methods we choose, our focus should deemphasize the technicalities of conducting statistical procedures, in favor of understanding their appropriate use and interpretation (Quilici & Mayer, 1996).

Thus, the lens through which middle-grades research in this study will be examined is 1) score reliability is reported for the data in hand (Capraro & Capraro, 2002; Henson, Capraro, & Capraro, 2002; Thompson, 2003); 2) null hypothesis statistical testing methods should be clear; 3) effect sizes (Capraro, 2004; Thompson 2007) and 4) confidence intervals (AERA, 2006; Capraro & Capraro, 2003) should always be reported.

## METHODS

For the purposes of this article three journals were considered that specifically report middle-grades research, *Middle Grades Research Journal*, *Research in Middle Level Education*, and *The Middle School Journal*. While *The Middle School Journal* deals mainly with the middle grades it does not contain original reports of empirical research and was eliminated from further consideration. Therefore, the remaining two journals met the criteria 1) dealing specifically with middle-grades research without limitation to any particular content area of developmental or structural aspects of middle grades, and 2) primarily contains reports of original empirical research. The inception of the *Middle Grades Research Journal* in Spring 2006 expanded opportunities for publishing middle-grades research presumably increasing the vitality of research in the area and providing additional signature space for the field. Therefore, for purposes of this study, the period from which articles were selected was from Spring 2006 through May 2009.

Article selection was done by reviewing each article in each issue for the given time period. For an article to be selected for coding it had to contain reports of quantitative analyses and sufficient details about the sample to make generalizability claims clear. Too few articles provided sufficient detail to classify articles by design.

### *Coding*

The coding criteria were based on recommendations contained in the guidelines established by the AERA Task Force (2006) and Wilkinson et al. (1999) but only a subset of those criteria were chosen for coding. The following categories were used: (a) descriptive statistics, (b) null hypothesis statistical significance testing (NHSST)/inferential statistics, (c) score reliability, (d) effect sizes, (e) confidence intervals (CI), and (f) types of statistical analyses. Each category contained additional indicators scored dichotomously as present or not present in each quantitative study. Among the indicators were types and quantity of descriptive statistics, NHSST, inferential statistics, types of score reliability and reliability induction, types and interpretation of effect sizes or contextualized effect, quantity and interpretation for CI, quantity and types of statistical analyses and if an alpha correction was employed when numerous univariate techniques were used, and finally quantity and types of graphical representations of the data. Of the 34 articles in MGRJ and the 32 articles in RMLE, 47 were quantitative, yielding 465 separate statistical analyses.

## RESULTS

The findings were reported as a composite of the two journals by coding criteria.

### *Descriptive Statistics*

Of the 38 studies reporting descriptive statistics (58% of the articles), many included standard deviations along with the means. In five cases, insufficient descriptive statistics were provided, some component, means, standard deviations, or a referent to what the reported descriptives were referring to was omitted. Descriptive statistics were less likely to be reported for the sample (58%) than for the variables of interest (73%). For example, it was common to get a mean for some measure but it was rarely disaggregated by subpopulation even in cases where the analyses focused on subsamples such as males and females within treatment or control, or ethnicity. Descriptive statistics were exclusively reported in the form of means and standard deviations but several studies did not include standard deviations for the means (25%). For those 25% of the studies, effect sizes for tests of differences between means were also not reported. Therefore, another researcher who may be interested in meta-analytically using those findings would have difficulty calculating the effect sizes because the exact *p*-value does not contain information about the obtained effect (for more information about effect sizes, consider Thompson and Zhang elsewhere in this volume).

### *Null Hypothesis Statistical Significance Testing (NHSST)/Inferential Statistics*

The NHSST and inferential statistics generally included at least one exact *p*-value (68%) but in five manuscripts authors who reported exact *p*-values also reported $p = 0.00$. This error results from rounding within statistical packages. If the researcher were to double click the value in SPSS, it would reveal a very small value represented in scientific notation. It should have been reported as less than .001 if indeed it was. Additionally, authors who reported exact *p*-values for *p*-calculated when it was less than .05 tended to not report exact *p*-values (80%) when it was greater. The inferential statistics were often underreported. It was common to invoke tests without examining the assumptions underlying their use. The three most abused were the failure to examine normality or departure from normality before choosing parametric or nonparametric statistics (93%), failure to discuss independence of observation before using ANOVA (100%), and using multiple univariate analyses without experiment-wise Type I error correction. It is important to note that the data are univariate normal supporting parametric statistics; however, when they were not univariate normal, nonparametric statistics should be invoked. Independence of observation is not about the data but about the study design. When the design is fully nested, such as students within teacher or teacher within school - one fails to meet this assumption. Just under 25% of the studies used any correction even though this is a very simple correction that can be applied to alpha or *p*-critical. The formula for a Bonferroni correction is $1 - (1 - \alpha)^{1/n}$ , where α equals the a priori alpha level, .05 in most cases, and n equals the number of comparisons. Therefore, in the study that reported 70 *t*-tests, a critical value of .00073 should have been used. This is a safeguard against multiple tests of statistical significance on the same data falsely giving the appearance of significance, because at $\alpha = .05$, 1 out of every 20 hypothesis-tests is expected to be statistically significant purely due to chance. The problem is that we do not know which one. Therefore, each test must be set to a substantially more stringent alpha level. A simpler, correction is to divide the a priori alpha level by the number of tests, therefore, .05/70 would equal an alpha level of .00071, essentially the same as the previous estimate. For the three studies reporting the most number of univariate tests, there would have been no statistically significant results to talk about. Additionally, for the next four highest, nearly half the results reported as statistically significant would not have been if they had used a corrected *p*-critical.

In addition, nearly 28% of the articles contained arbitrary *p*-critical values, for example in some places $p < .70$ was reported, on occasion within the same paragraph and set of analyses a $p < .031$ and $p <. 003$, were used without a rationale. It is important to set the *p*-critical value before one begins analysis of the data and that criteria should be stated in the article along with reporting the exact *p*-calculated value to three places.

### *Score Reliability*

It is important to report the reliability for the data in hand because it attenuates the obtained effects. Twenty of the studies reported a Cronbach's alpha reliability and one reported Cohen's Kappa for the data in hand. Unfortunately, 14 reported a reliability coefficient from a test manual or source other than their study or reliability induction. The obtained reliability may be very different from that reported by other researchers or from the test manual. Because researchers often perform statistical analyses that consider subsamples of the dataset, reliability should be reported for all subsamples of interest because it is likely the reliability for a subgroup can vary greatly from the entire sample. Only two studies considered all the reliabilities for the subsamples of interest that were used in their analyses. Only two articles included all the items for all the instruments used in the study. It is difficult to accumulate knowledge when a researcher-designed instrument is used once and not provided for subsequent authors to use in subsequent studies. While it is acceptable practice for authors to say the instrument is available by request, academia affords a great deal of mobility and tracking down authors can be daunting. Additionally, it is possible that a particular line of research would not achieve a following until an author is no longer in academia and the instrument would be lost to future generations of researchers attempting to further that particular line of inquiry.

### *Effect Size*

Nearly 70% of the studies reported some estimate of effects. The most common metric was percent change (48%) of measured variables followed by $eta^2$ and ANOVA, however, $eta^2$ was not reported for all ANOVAs. A phi coefficient, reported with non-parametric methods was only reported 6% of the time while non-parametric methods were represented quite heavily. Typically, regression results include an $R^2$ effect, however, in these two journals it was only reported 21% of the time (see discussion on regression in Yetkiner elsewhere in this volume). Cohen's *d* was used 15% percent of the time but almost never with *t*-tests. The Cohen's *d* is among a class of effect sizes that can be easily computed. The formula is $d = \frac{\overline{X}_E - \overline{X}_C}{\sigma}$, the difference in the means divided by the standard deviation for the whole sample. Even though this formula is straight forward it was underused given *t*-test were so heavily represented in the articles.

### *Confidence Intervals*

Thinking meta-analytically is important and CIs function to provide more information than a single point estimate. CIs can provide a graphical representation of the preciseness of the estimate as well comparative estimates when considering other similar studies. Only one article reported and graphed CIs. For more information about CI's see Skidmore elsewhere in this volume.

### *Types of Statistical Analyses*

The most common analytic choice by number of times it was used was the *t*-test ($n$ = 212; range of 1 - 70 times per article); followed by $\chi^2$ ($n$ = 120; range 1-102); ANOVA ($n$ = 65; range 1-10); regression ($n$

=19; range 1 - 3); and correlations ($n$ = 15) (see Zhang for more information about $t$-tests and Yetkiner for a multivariate option elsewhere in this volume). If each correlation had been counted individually, it would have been the most widely used analytic choice. In some articles, correlation coefficients were computed for all variables and latent variables by sub-groups, even though none of the coefficients were interpreted. Some complex analyses were also reported albeit to a much lesser extent. Hierarchical regression was used 10 times but in only two articles. Exploratory and confirmatory factor analyses were used a total of 6 times in three articles. MANOVA represented about 3% of the analyses across the articles. Using MANOVA reduces the number of univariate analyses thus managing Type I error.

## FOR THE MIDDLE-GRADES RESEARCHER, CONSUMER, & REVIEWERS

The important messages for those who write for the middle-grades community is to make a cognizant effort to adhere to recommendations that will improve the quality of what we know and how we can best situate our research reports to provide maximum information for the reader. The typical quantitative research article should provide the psychometrics for the instruments being used, minimally the reliability for the data in hand, plus (if applicable) a copy of the instrument in the appendix or provide availability from the author. Next, it is important to inform the reader of how null hypothesis testing will be evaluated making use of corrected $p$-values when planning to use multiple univariate tests. It is also important to make explicit how implicit assumptions for each analytic choice is met. Because small sample sizes limit the power to reject the null hypothesis – it is always a good idea to report the effect size for every hypothesis test. Some may hold that unless the results are statistically significant, one does not need to consider the magnitude of the effect or that a statistically significant result automatically makes the obtained effect practically important. Both of these are false conceptions. First, a small sample size can mask a practically important effect and this information may have real implications for theory development and should not be brushed aside because it was masked by a small sample size. Secondly, if one has a sample of 2000 participants, a very small difference will be statistically significant even when that difference is not practically important. Therefore, obtaining statistically significant results do not assure practically important differences.

One option for reducing the reliance on $t$-test and chi square analyses is to use multivariate methods that are more effective (consider Yetkiner elsewhere in this volume). Using MANOVA allows two or more dependent variables to be compared to several independent variables which can provide a much more parsimonious model and a better portrayal of reality.

Confidence intervals provide insights into the precision of the estimate as well as a metric for comparing variables. Additionally, it helps to conceptualize that the point estimate obtained in every study is just an estimate and many other estimates are equally plausible.

We offer a set of guidelines for middle-grades researchers who engage in quantitative research, rooted in an evaluation of practice for consideration when writing reports of original research to maximize the transportability of their work and to facilitate its usefulness. These topics are no less important, but the What Works Clearing House has already emphasized issues related to design and participant selection so reemphasizing them here would limit the message rooted in the reporting practice of middle- grades researchers. We applaud the tireless work of those who participate in the peer-review process. Without their efforts and the dedication of editors, reviewers, and authors this manuscript would not have been possible nor would the field have the opportunity for reflecting and reexamining its cumulative work.

The four-part lens of score reliability, null hypothesis statistical significance testing, effect sizes, and confidence intervals, through which middle grades research was examined yielded recommendations that these authors hope middle-grades journals, editors, reviewers and authors will embrace when using quantitative methods:

1. requiring the instrument in initial publications and psychometrics each time it is used
2. reporting alpha a priori and any requisite corrections when using multiple univariate tests
3. reporting exact *p*-values to three decimal places and less than .001 when smaller thus never reporting a calculated *p*-value as .000
4. reporting and discussing effect size estimates and making explicit comparisons to those reported in other similar studies
5. limiting the use of multiple univariate tests and increasing the use of multivariate tests

**REFERENCES**

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

Capraro, R. M. (2004). Statistical significance, effect size reporting, and confidence intervals: Best reporting strategies. *Journal for Research in Mathematics Education, 35*, 57-62.

Capraro, M. M., & Capraro, R. M. (2003). Exploring the impact of the new APA 5th Edition Publication Manual on the preferences of journal board members. *Educational and Psychological Measurement, 63*, 554-565.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher,31*(8), 4–14.

Henson, R., Capraro, R. M., & Capraro, M. M. (2004). Reporting practice and use of exploratory factor analysis in educational research journals. *Research in the Schools, 11*(2), 61-72.

Natriello, G. (2006). Research and educational policy (Part 1): The relationship between research and policy. *Teachers College Record.* Retrieved July 1, 2009 from http://www.tcrecord.org, ID No. 10636

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 Press.

Stipek, D. (2005). 'Scientifically based practice': It's about more than improving the quality of research. *Education Week, 24*(28), 34-35.

Thompson, B. (Ed.) (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

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.

Wilkinson, L., & APA 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*(24), 38-39.