

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to the *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 16, Number 14, October 2011

ISSN 1531-7714

Reporting and Discussing Effect Size: Still the Road Less Traveled?

James H. McMillan

Jennifer Foley

Virginia Commonwealth University

This study shows the extent to which effect size is reported and discussed in four major journals. A series of judgments about different aspects of effect size were conducted for 417 articles from four journals. Results suggest that while the reporting of simple effect size indices is more prevalent, substantive discussions of the meaning of effect size is lacking.

Despite the admonitions of many statisticians and educational researchers (e.g., Fidler and Cummings, 2008; Thompson, 2008), and newer AERA standards “requiring” effect sizes to be reported and interpreted for every essential statistical result (AERA, 2006), universal reporting and discussing effect size in quantitative studies remains elusive. Using specific effect size statistics, or even the concept of magnitude of findings as different from statistical significance, is clearly not yet integral to conducting and reporting educational research. This appears to be true even though effect size has been addressed in most statistical methods textbooks for three decades (Huberty, 2001), well-known and respected methodologists have written about effect size in leading journals and in books (e.g., Cohen, 1994; Ellis, 2010; Glass, 1976; Grissom & Kim, 2005; Kirk, 1996; Rosenthal, 1991; Rosenthal & Rubin, 1982; Wilkinson & APA Task Force on Statistical Inference, 1999; Thompson, 2007; and Vacha-Haase & Thompson, 2004), and over 23 journals have adopted editorial policies that require effect size reporting (Vacha-Haase & Thompson, 2004).

This article will further educate researchers on the importance of substantive reporting of effect size. The purpose of this study is to partially replicate an earlier investigation of effect size reporting in four well known educational research journals. The study includes a determination of the

specific nature of effect size indices reported, as well as the extent to which researchers discuss magnitude of effect in analyzing their findings, and compares these findings to earlier research to get some idea of the impact of increasing attention to effect size in the last decade.

Review of Related Literature

Eleven previous reviews of different educational and psychological journals published from 1990 through 2002 indicate considerable variability between journals and little increase over time in the percentages of articles that included effect size indices (Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000). The typical methodology employed in these studies was to review every article in all journal issues of a designated volume. Only two of the studies reviewed four or more journals (Keselman, Huberty, Lix, Olejnik, Cribbie, Donahue, Kowalchuk, Lowman, Petoskey, Keselman, & Levin, 1998; and Kirk, 1996), and a number of important journals for educational research were either omitted or reviewed for only one or two years. Furthermore, many of the studies were simple counts of the frequency of effect size indices used (Henson & Smith, 2000). A more meaningful review, included with the present study, indicates, as suggested by Vacha-Haase & Thompson (2004), whether there was any discussion or interpretation of effect size. This is important since it is necessary to both

interpret and evaluate effect size (Kline, 2004), not to simply include the statistical measure of effect size without discussion.

Four more recent studies, one in psychology, two in different areas of education, and one in both education and psychology, also show the extent and nature of use of effect size. Sun, Pan, and Wang (2010) reviewed effect size reporting from 1,243 articles published in 14 journals, from 2005–2007 (Table 1).

Table 1: Summary of Effect Size Reporting in 14 Educational and Psychological Journals, 2005–2007¹

Category	Total <i>n</i>	Reported ²		Interpreted ³	
		<i>n</i>	%	<i>n</i>	%
Journal type					
AERA ⁴	69	50	73	31	62
APA ⁵	863	349	40	179	51
Independent ⁶	311	211	68	136	64
Total	1,243	610	49	346	57
Year					
2005	438	198	45	116	59
2006	422	207	49	115	56
2007	383	205	53	115	56

¹Sun, Pan, & Wang (2010).

²Included Cohen’s *d*, *f*, and *b*; *r*, R^2 ; η^2 and partial η^2 ; odds ratio.

³Percentage based on number reported.

⁴*American Educational Research Journal, Educational Evaluation and Policy Analysis.*

⁵*Journal of Educational Psychology, Journal of Experimental Psychology: Applied, Journal of Experimental Psychology: Human Perception and Performance, Journal of Experimental Psychology: Learning, Memory, and Cognition, School Psychology Quarterly, Dreaming.*

⁶*Early Childhood Research Quarterly, Journal of Experimental Education, Journal of Learning Disabilities, Learning Disabilities Research & Practice, Journal of Special Education, Infant and Child Development.*

They found that up to 49% of the articles reported effect size, including about 70% of articles in two AERA journals (*Educational Evaluation and Policy Analysis* and *American Educational Research Journal*), though only slightly more than half of these articles analyzed and/or interpreted the effect size results. They also found that the most frequently used effect size measure was some variation of strength of relationship, often R^2 . The authors conclude that

the overall rate of reporting effect size “is still far from satisfactory” (p. 998).

Alhija and Levy (2009) conducted a content analysis of 183 statistical analyses reported in 99 randomly selected education articles from both educational and psychological journals (10 total journals). Examining two years of articles, 2003–2004, they found that most studies investigating strength of relations reported effect size (86%, including *r*, R^2 , and chi-square), while a much smaller percentage of studies examining differences reported effect size (57%, including Cohen’s *d*, η^2 , and partial η^2). They also found only a small difference between journals “requiring” effect size compared to journals “not requiring” effect size, with slightly more for journals with a policy of requiring effect size (Table 2). However, they included *r* as an effect size indicator. Though the actual number of analyses that utilized *r* is not reported separately, the inclusion of it would inflate the number of effect sizes. Also, consistent with Sun et al, there was little interpretation or discussion of effect size in these studies.

Table 2: Summary of Effect Size Reporting for Quantitative Analyses in 99 Randomly Selected Articles from 10 Educational Journals, 2003–2004^{1,2,3}

Category	Total <i>n</i>	Reported	
		<i>n</i>	%
Journal type			
“Requiring” effect size	100	74	74
“Not requiring” effect size	83	56	67
Measure type			
Mean differences ⁴	99	56	57
Strength of relations ⁵	74	64	86

¹Alhija & Levy (2008)

²Included Cohen’s *d* and *f*; *r*, R^2 , β , and “variance explained;” η^2 and partial η^2 ; chi-square.

³*Contemporary Educational Psychology, Journal of Learning Disabilities, Exceptional Children, Early Childhood Research Quarterly, Journal of Experimental Education, Journal of Educational Psychology, Learning Disabilities Research & Practice, Journal of Special Education, Infant and Child Development, and Educational Research.*

⁴Includes Cohen’s *d*, η^2 , and partial η^2 .

⁵Includes *r*, R^2 , and chi-square.

Zientek, Capraro, and Capraro (2008), analyzed 217 statistical tests from a review of 174 quantitative articles concerning evidence of best practices in teacher education published prior to 2005. Effect sizes were found in 39% of the cited articles. However, they included Pearson r and Spearman ρ as effect size indices. Not surprisingly, they found that 94% of studies reporting correlations or regressions included effect size, and only 7% of studies using difference effect size statistics, such as Cohen's d . They also found little discussion of effect sizes when reported, mostly about results of zero-order correlations and regression analyses. Similar results were found by Mathews, Gentry, McCoach, Worrell, Matthews, and Dixon (2008), in a study of 101 quantitative journal articles investigating gifted education, using articles from five journals across ten years (1996-2006). They found that the percentage of articles reporting effect size increased from 26% during 1996-2000 to 46% during 2001-2005. Approximately 65% of the indicators were for correlational findings, which included reporting zero-order correlation coefficients. As with other previous reviews, this suggests reports of high percentages of articles using effect size were spurious.

Previous investigations of effect size reporting suggest that, despite recommendations in the 1994, 2001, and 2010 editions of the APA *Publication Manual* (e.g., "For the reader to appreciate the magnitude or importance of a study's findings it is almost always necessary to include some measure of effect size in the Results section," APA *Publication Manual*, 2010, p. 34) and of AERA (AERA, 2006) that essentially admonish researchers to use effect size information, many researchers are still not including effect size indicators. The McMillan, Lawson, Lewis, and Snyder (2002) study found that, of the 508 articles that were either quantitative, mixed-method, or simulation, 148 (29%) at least mentioned or calculated an effect size. Only 82 of these 508 articles (16%) included both a calculation of effect size and at least limited discussion of magnitude or practical significance. Only 30 of 508 articles (6%) included both a calculated effect size and what was judged to be extensive discussion (typically several sentences or more of interpretation

of magnitude or practical significance). In particular, Thompson (2008) argues that greater discussion of effect size results is needed. Of interest with the present study is to investigate the progress of our profession in following repeated recommendations by experts, journals, and professional associations for both reporting and discussing the nature of effect size.

Research Questions

The purpose of this study is to partially replicate the McMillan et al. study of effect size reporting in four well known educational research journals. The investigation includes a determination of the specific nature of effect size indices reported as well as the extent to which researchers discuss magnitude of effect in analyzing their findings. The study then compares these findings to earlier research to get some idea of the impact of increasing attention to effect size in the last decade. More specific research questions included:

1. What percentage of quantitative or mixed-method empirical studies reported and discussed effect size?
2. What specific effect size estimates are used?
3. What is the extent of interpretation and discussion of the meaning of the effect size estimates?
4. How do recently published studies' reporting and use of effect size compare with other investigations of effect size?
5. What is the implication of the current nature of effect size reporting for graduate training and professional development of educational researchers?

Data Source

Quantitative and mixed-method empirical studies were reviewed in four journals (*Journal of Educational Psychology*, *Journal of Experimental Education*, *Journal of Educational Research*, and *Contemporary Educational Psychology*), from 2008-2010. The selection of journals was made to provide a

representative sample that would present a reasonable indication of the extent to which educational researchers used effect size indices. It included journals that are well known, widely distributed, and publish a relatively large number of quantitative studies. These same four journals were reviewed previously by McMillan, et al.

Method

A systematic content analysis of each article was conducted, following closely earlier reviews by Sun et al., and Alhija & Levy. A Data Extraction Form (Appendix B) was pilot tested and reviewed by experts¹ and revised to enable objective recording of rater judgments (see Appendix A). For each quantitative article there was a determination of whether any effect size indicators were reported, and the level of discussion of magnitude based on these findings. For articles that included an effect size index, the specific type of index was determined under two general categories - difference and relationship - with note of the specific reported statistic (e.g., Cohen's *d*, *r*², *eta*²). Discussion and interpretation was rated according to suggestions in the literature, such as comparing effect sizes to those from previous research, relating to sample size, considerations of the context of the study, and extensiveness. Examples of discussion and interpretation, when present, were recorded.

Interrater reliability was established on a random selection of 61 journal articles (15% random sample). For each article there were 17 judgments, resulting in a total of 1,037 judgments. Interrater agreement was found on 936 categories (90%). For the single item that used a scale rather than dichotomous category, agreement within 1 point on the scale was 87%.

Results

Table 3 presents frequencies and percentages of articles in each of the four journals by dominant methodology. A total 417 articles were reviewed. As expected, there was a high percentage of

quantitative and mixed-method studies reported (92%). This included 98% of the articles in the *Journal of Educational Psychology*. Only the *Journal of Educational Research* included a number of qualitative articles during the three years (12%). It is important to note that the *Journal of Educational Psychology* had considerably more articles than the other three journals, which results in a large portion of the total (47%).

Table 3: Number and Percent of Journal Article Methodologies by Journal and Year

Category	Tot <i>n</i>		Quanti- tative		Mixed- Method		Quali- tative		Other	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Journal										
1	86	75	87	4	5	1	1	6	7	
2	198	193	98	1	1	2	1	2	1	
3	83	52	63	11	13	10	12	10	12	
4	50	45	90	2	4	2	4	1	2	
Year										
2008	148	126	85	11	7	8	5	3	2	
2009	151	136	90	3	2	4	3	8	5	
2010	118	103	87	4	3	3	3	8	7	
Total	417	365	88	18	4	15	4	19	5	

Journals: 1. *Contemporary Educational Psychology*, 2. *Journal of Educational Psychology*, 3. *Journal of Educational Research*, 4. *Journal of Experimental Education*

Table 4 shows the frequencies and percentages of quantitative and mixed-method studies that reported some type of effect size indicator. Frequently, more than one index was reported in the same article. For all studies, 74% used a specific indicator of effect size. The percentages ranged from a high of 82% for the *Journal of Educational Research* to a low of 64% for the *Journal of Experimental Education*. There was no clear trend across the three years. Three effect size indicators dominated those reported (Cohen's *d*, *eta*² and partial *eta*², and proportion of variance accounted for [*R*²]). There was very little use of Hedges's *g*, odds ratio, Cohen's *f*, or *omega*². Authors publishing in *Journal of Educational Psychology* used proportion of variance and Cohen's *d* more than the other three journals.

Table 4: Number and Percent of Effect Size Indicators Reported in Quantitative and Mixed-Method Articles by Journal and Year

Category	Reported		Cohen's <i>d</i>		Hedges's <i>g</i>		Eta ² ; partial Eta ²		Odds Ratio		R ² ; proportion of variance		Omega ²	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Journal														
1	42	70	17	23	0	0	21	28	1	1	18	24	1	1
2	143	76	60	32	1	1	37	20	6	3	81	44	1	1
3	50	82	16	26	1	1	19	31	7	12	23	38	1	2
4	30	64	10	21	0	0	13	28	3	6	14	30	0	0
Year														
2008	94	72	37	29	0	0	33	26	4	3	48	37	0	0
2009	104	77	36	27	1	1	38	28	5	4	49	37	1	1
2010	67	74	30	29	1	1	19	18	8	8	39	37	2	1
Total	265	74	103	28	2	1	90	25	17	5	136	37	3	

Journals: 1. *Contemporary Educational Psychology*, 2. *Journal of Educational Psychology*, 3. *Journal of Educational Research*, 4. *Journal of Experimental Education*

Our interpretation of how the authors discussed effect size indicators that were reported is summarized in Table 5. When Cohen's *d* was reported it was typically followed by an indication that Cohen's guidelines (Cohen, 1988) for interpretation were used (.2= "small," .5= "medium," and .8= "large"). The percentage of articles using this convention, as a percentage of articles reporting Cohen's *d*, was very high (94%). Few articles mentioned the What Works Clearinghouse guidelines for interpretation of the

magnitude of effect expressed as a function of standard deviation, like Cohen's *d*. There was comparatively little discussion of the meaning of the effect size results, whether as compared to effect sizes in previous research, the context of the study, or as related to sample size. This was true for each of the journals without clear patterns of differences across years.

Table 6 presents the results for the single item that shows the judgments of the raters concerning the extensiveness of the interpretation and/or

Table 5: Number and Percent of Nature of Interpretation and Discussion of Effect Size Indicators Reported in Quantitative and Mixed-Method Studies by Journal and Year

Category	Cohen's <i>d</i> guidelines		WWC guidelines		Compared to previous research effect size		Considers context		Explicitly related to sample size	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Journal										
1	20	26	0	0	7	9	9	12	5	7
2	56	30	0	0	33	18	29	15	15	8
3	13	21	2	3	8	13	5	8	4	7
4	8	17	0	0	4	9	9	19	4	9
Year										
2008	30	23	0	0	15	12	17	13	8	6
2009	36	27	1	1	22	16	17	13	13	10
2010	31	29	1	1	15	14	18	17	7	7
Total	97	26	2	1	52	14	52	14	28	8

Journals: 1. *Contemporary Educational Psychology*, 2. *Journal of Educational Psychology*, 3. *Journal of Educational Research*, 4. *Journal of Experimental Education*

discussion of effect size. Appendix A contains excerpts from articles to show differences between different levels of interpretation. Approximately half of the articles did not include any discussion or interpretation of effect size results. That is, the authors of these studies typically showed Cohen’s d , η^2 or R^2 as part of reporting statistical significance, without any interpretation. This was the case for all four journals. Nearly 20% of the articles included “very extensive” or “extensive” interpretation and discussion, amounting to three or more sentences.

size, at least in these four journals, weighted heavily by the *Journal of Educational Psychology*.

Discussion

The first important issue with this study, within the context of other systematic reviews of articles for effect size, is the determination of what does and does not “count” as an effect size indicator. In our study, unlike most others, we chose not to include the reporting of r and chi-square as effect

Table 6: Extensiveness of Interpretation and Discussion of Effect Size Indicators Reported in Quantitative and Mixed-Method Studies by Journal and Year¹

Category	Very Extensive		Extensive		Some		Little		None	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Journal										
1	5	7	6	8	8	11	14	19	41	55
2	24	13	18	10	36	19	18	10	92	49
3	7	12	18	10	36	19	16	26	23	38
4	4	9	3	6	5	11	9	19	26	55
Total	40	11	31	8	60	16	57	15	182	49

¹Very extensive – more than 5 sentences; Extensive – 3 to 5 sentences; Some – 2 or 3 sentences; Little – one sentence

The final data from the review, summarized in Table 7, shows the percentages of quantitative and mixed method studies reporting effect size indicators in the same four journals from 2008-2010 compared to articles published from 1997-2000. The percentage changes show substantial increases in reporting effect size indices over this approximate ten-year period. All four journals showed increases,

size statistics. This may explain why our findings showed a much lower percentage of interpretation as compared to the reviews by Sun et al. (2010) and Alhija and Levy (2008). It seems that there is a need for further clarification of what constitutes effect size and why it is important.

Table 7: Percentage of Quantitative and Mixed-Method Studies Reporting and Interpreting Effect Size by Year

Category	Reported			Interpreted		
	1997-2000 ¹	2008-2010	Change	1997-2000 ¹	2008-2010	Change
Journal	%	%	%	%	%	%
1	35	70	+ 35	39	45	+ 6
2	23	76	+ 43	23	51	+ 28
3	35	82	+ 47	34	62	+ 28
4	44	64	+ 20	44	45	+ 1
Total	30	74	+ 44	30	51	+ 21

¹McMillan, Lawson, Lewis, Snyder (2001)

²Little, Some, Extensive, or Extensive

with an overall increase of 44%. This suggests that many more researchers are now reporting effect

In many of the studies we reviewed, Cohen’s d and η^2 were simply listed along with indicators of statistical significance, without further reference to it, as if d and η^2 are meaningful simply by being reported. This interpretation is consistent with Zientek et al. (2008) and Kirk (1996), who also found reporting with little if any discussion. This suggests that it is difficult to know if the researchers actually understood effect size or merely included an effect size statistic.

It seems to us that zero-order correlations, by themselves, should not be interpreted as an effect size indicator, despite Cohen’s inclusion of it in his list of effect size statistics. The proportion of variance accounted, with the coefficient of determination or coefficient of multiple determination, is a much better indicator of effect

size. Of course you can mentally square the correlation easily enough, but without including this squared statistic the meaning of proportion of variance accounted for is unclear. R^2 , on the other hand, which is commonly reported in studies using regression, does indicate proportion of variance and is a reasonable measure of effect size. With R^2 there is an understanding of how much variability is and is not accounted for, which helps in interpreting the extent to which a relationship is a function of measured variables.

Considering that we did not include r as a measure of effect size, our finding that 64% - 82% of the quantitative and mixed method articles at least included a measure of effect size is encouraging. Our percentages reflecting use were somewhat higher than what was reported by Alhija & Levy (2008), and Sun et al. (2010). This may be because of differences in methodology, journals used for the review, and time frame, which reflected findings in earlier years. A direct comparison is not possible. Our findings clearly indicated an increase in use of effect size from 1997 to 2010, though this also may be partially a function of different raters and differences in specific rating forms. Only one rater was the same for both analyses.

The extent of interpretation also appears to have increased from 1997 to 2010 for the four journals in this study, albeit to a lesser extent than reporting. In the earlier study the rating of interpretation was based on a four point scale, ranging from none to 3, with 3 indicating "extensive" interpretation. The current study used a five point scale and attempted to capture discussion with a more concrete indicator. While the number of sentences is not the same as the more qualitative judgment used in the earlier study, the 19 percentage rated "very extensive" or "extensive" in the current study is about the same as the 6% rated "extensive" (3) or 23% rated as a 2 or 3 in the earlier study of articles from the same journals. This consistently low percentage shows little progress toward interpretation of effect size. Our current data show much more interpretation with Cohen's d and R^2 than other indicators, with lower percentages explicitly discussing effect size in relation to the size

of the sample, context, or effect size from previous studies. This is not surprising, given the ubiquitous though often thoughtless use of Cohen's guidelines, and the use of R^2 for many years with interpretation of regression analyses.

Sun et al. reported a higher percentage of articles that interpreted effect size. In their study there are no examples of what they considered "interpreted," which, along with the inclusion of different journals and raters, could explain why their percentage was so much higher than what we found. When results of the current study considered interpretation only for quantitative and mixed-method studies reporting effect size, 46% were rated as having at least "some" discussion, still lower than the Sun et al. findings.

It was interesting to find only six articles that used guidelines from the WWC. This is important since the WWC has made an extensive effort in suggesting that an effect size that shows only a "small" effect, as per Cohen, is probably meaningful and important for practice. That is, a difference of only a quarter or fifth of a standard deviation in educational studies is considered important in many contexts. This highlights the current limited interpretation and use of effect size. As many have argued, the reason we have effect size is to show the magnitude and importance of the results, explained in such a way so that it is clear what the results mean with respect to original units or units that are relevant to understanding the practical impact. Effect size transforms abstract statistical significance testing to concrete measures of relationship or difference. This is what the WWC is trying to show in their evaluation of reported effect sizes. Researchers publishing in the four journals reviewed for this study may not be aware of the WWC guidelines, and/or may not understand the importance of discussing what effect size results mean. Judging from the low percentage even interpreting effect size results, there appears to be a need for researchers to emphasize, analyze, and report the meaning of results in units and in ways that give others an indication of the magnitude and importance of the findings.

In summary, our findings suggest that when it comes to effect size the number of researchers reporting and interpreting such statistics is improving. While the number of studies using effect size has increased, analysis of the nature of the effect size and meaningfulness of the results is lacking. Many researchers are using effect size without sufficient discussion of what it actually suggests in the context of their study and for practice. Furthermore, there was reliance on Cohen's *d* guidelines for interpreting effect size results, rather than what is recommended and used by the What Works Clearinghouse, which were designed with the educational setting in mind. This suggests a need to clarify for researchers which guidelines make most sense within the context of their research, and what considerations should be analyzed to discuss and interpret effect size results, consistent with recommendations from Thompson (2008). While difference indices are appropriate for many studies, more attention may need to be given to relationship indicators.

The trends reported here, albeit with only four journals, have demonstrated typical usage of effect size in educational research and suggest that there is need for researchers and statisticians to make more effective use of effect size in reporting results and discussion of practical significance. Those teaching research have a specific responsibility to educate students about the nature and use of effect size so that meaningfulness and practical importance are clearly understood.

References

- Alhija, F. N.-A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement, 69*(2), 245-265.
- 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. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.
- Byrnes, J. P., & Wasik, B. A. (2009). Factors predictive of mathematics achievement in kindergarten, first and third grades: An opportunity-propensity analysis. *Contemporary Educational Psychology, 34*(2), 167-183.
- Cohen, J. (1994). *Statistical power analysis for the behavioral sciences* (2ed). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Daniels, L. M., Haynes, T. L., Stupinsky, R. H., Perry, R. P.; Newall, N. E., & Pekrun, R. (2008). Individual differences in achievement goals: A longitudinal study of cognitive, emotional, and achievement outcomes. *Contemporary Educational Psychology, 33*(4) 584-608.
- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press.
- Fidler, F., & Cumming, G. (2008). The new stats: Attitudes for the twenty-first century. In J.W. Osborne (Ed.), *Best practices in quantitative methods*. Thousand Oaks, CA: Sage Publishing.
- Gehlbach, H., Brown, S. W., Ioannou, A. Boyer, M.A., Hudson, N., Niv-Solomon, A., Maneggia, D., & Janik, L. (2008). Increasing interest in social studies: Social perspective taking and self-efficacy in stimulating simulations. *Contemporary Educational Psychology, 33*(4), 894-914.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.
- Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.
- Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA Task force report and current trends. *Journal of Research and Development in Education, 33*, 285-296.
- Huberty, C. J. (2001). A history of effect size indexes. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA Analyses. *Review of Educational Research 68*(3), 350-386.

- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.
- Klassen, R. M. Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology, 102*(3), 741-756.
- Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.
- Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2008) A multilevel study of predictors of student perceptions of school climate: The effect of classroom-level factors. *Journal of Educational Psychology, 101*(1), 96-104.
- Marsh, H., Martin, A., & Cheng, J.H.S. (2008). A multilevel perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology, 100*(1), 78 –95.
- McMillan, J. H., Lawson, S., Lewis, K., & Snyder, A. (2002, April). Reporting effect size: The road less traveled. Paper presented at the 2002 Annual Meeting of the American Educational Research Association, New Orleans.
- Muis, K. R., & Franco, G. M. (2009) Epistemic beliefs: Setting the standards for self-regulated learning, *Contemporary Educational Psychology, 34*(4), 306-318.
- Nelson, R. M., & DeBacker, T. K. (2008). Achievement motivation in adolescents: The role of peer climate and best friends, *Journal of Experimental Education, 76*(2), 170 – 189.
- Pekrun, R., Goetz, T., Daniels, Lia M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion.” *Journal of Educational Psychology, 102* (3), 531-549
- Powell, D. R., Diamond, K. E., Burchinal, M.R., & Koehler, M. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of Educational Psychology, 102*(2), 299-312.
- Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indexes. *American Psychologist, 46*, 1086-1087.
- Rosenthal, R., & Rubin, D.B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.
- Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102*(4), 989-1004.
- Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423-432.
- Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 246-262). Newbury Park, CA: Sage.
- Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10*(3), 413-425.
- Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*, 473– 481.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.
- 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.

Appendices

Appendix A: Excerpts From Articles Showing Different Levels of Interpretation

Limited	Some	Extensive	Very Extensive
<ul style="list-style-type: none"> • Black boys had 55% greater odds of receiving a teacher-reported ODR compared with white boys. (Bradshaw, Mitchell, O' Brennan, & Leaf, 2010, p. 512). • Note how the effect sizes for ethnicity were considerably smaller than those for SES (Byrnes, & Wasik, 2009, p. 177). • Given that effect sizes ranged from moderate to large when changes occur, these changes were noteworthy (Muis, & Franco, 2009, p. 269). • The overall variances accounted for in the outer variables were modest, yet small effects might have larger accumulations effects over time (Nelson, & DeBacker, 2008, p. .) • Workload stress accounted for 31% of the variance in teachers' overall teaching stress Klassen, & Chiu, 2010, p. 747). 	<ul style="list-style-type: none"> • Finally, although some of our effect sizes are small according to conventional standards, they indicate systematic involvement of goals across three domains of academic outcomes. Given that we controlled for strong background variables including age and high school average, all additional variance explained in the dependent variables can be viewed as meaningful. (Daniels, Haynes, Stupinsky, Perry, Newall, & Pekrun, 2008, p. 604). • However, effect sizes in these differences were small ($d_s = -.28$ and $.27$ for control and elaboration...as in study 2, effect sizes were small ($d_s = .21$, $.23$, and $.29$) Pekrun, Goetz, Daniels, Stupnisky, & Perry, 2010, p. 538). 	<ul style="list-style-type: none"> • Specifically, children in intervention classrooms had significant gains in letter knowledge ($d = .29$) concepts about print ($d = .22$), writing ($d = .17$) and blending ($d = .18$).... • effect size results from prior studies of similar interventions and target populations provide a benchmark for interpreting the current study's effect sizes for significant intervention efforts. (Powell, Diamond, Burchinal, & Koehler, 2010, p. 307, 309). • Only 2 of 23 effects were statistically significant, and each of these was small. Girls had higher self efficacies (averages across the three subjects), but the difference was not large (approximately $.09$ standard deviations, as all the variables were standardized to facilitate interpretations) (Marsh, Martin, & Cheng, 2008, p. 85). 	<ul style="list-style-type: none"> • The proportion of variance suggests there is greater variability in students' willingness to learn within schools and that this aspect of school climate may be more indicative of individuals' own motivation than is overall aggregated perception. In contrast, the amount of school-level variance for order and discipline was much higher (27 %), suggesting that perceptions of school safety may be more relevant to school characteristics than achievement motivation; however, the individual level still accounted for the majority of the variance. Last, 8% to 9% of the variance across the two climate outcomes was attributable to clustering at the classroom level. This partitioning of variance is relatively consistent with previous research by Vieno et al. (2005), who found that 84% of the variation in climate was accounted for at the individual level, whereas 11% was accounted for at the class level, and just 4% at the school level. (Koth, Bradshaw, & Leaf, 2008, p. 101).

Appendix B: Data Extraction Form

Effect Size Reporting and Interpretation

Coder: _____

Date: _____

Journal: _____

- 1- *Contemporary Educational Psychology*
- 2- *Journal of Educational Psychology*
- 3- *Journal of Educational Research*
- 4- *Journal of Experimental Education*

Last Name of First Author: _____

Year: _____ Volume: _____ Number: _____

Dominant Nature of Study: _____

- 1- Quantitative
- 2- Mixed Method
- 3- Qualitative
- 4- Other, e.g., review of literature

For Quantitative and Mixed Method \rightarrow Research Design: _____

- 1- Experimental
- 2- Single Subject
- 3- Causal Comparative/Ex post facto
- 4- Correlational/Comparative
- 5- Other, eg., instrument development

Effect Size Index(s) Reported:

Specific index reported	Yes _____	No _____
Cohen's <i>d</i>	Yes _____	No _____
Hedges's <i>g</i>	Yes _____	No _____
Eta ²	Yes _____	No _____
Odds Ratio	Yes _____	No _____
Proportion of variance explained (<i>R</i> ² or <i>r</i> ²)	Yes _____	No _____
Omega ²	Yes _____	No _____

Other: _____

Effect size confidence intervals reported? Yes _____ no _____

Interpretation and Discussion:

Cohen's <i>d</i> Guidelines	Yes _____	No _____
WWC Guidelines	Yes _____	No _____
Comparison with effect size of previous research?	Yes _____	No _____
Includes consideration of context – what outcome is measured?	Yes _____	No _____
Explicitly related to sample size?	Yes _____	No _____

Extensiveness of Interpretation and Discussion: _____

1. Very extensive – more than five sentences
2. Extensive – three to five sentences
3. Some – two to three sentences
4. Little – one sentence
5. None

Nature of Interpretation and Discussion:

Other Comments:

Citation:

McMillan, James H. & Jennifer Foley (2011). Reporting and Discussing Effect Size: Still the Road Less Traveled? *Practical Assessment, Research & Evaluation*, 16(14). Available online: <http://pareonline.net/getvn.asp?v=16&n=14>

Acknowledgement

Reviewers of a draft of the Data Extraction Form were Gene Glass, Bruce Thompson, and Bob Slavin. Their help is appreciated.

Authors:

James H. McMillan
School of Education
Virginia Commonwealth University
Box 842020
Richmond, Virginia 23284-2020
Jmcmillan [at] vcu.edu

Jennifer Foley
School of Education
Virginia Commonwealth University
Box 842020
Richmond, Virginia 23284-2020