Available online at www.sciencedirect.com

ScienceDirect

Currents in Pharmacy Teaching and Learning 8 (2016) 83–89

Currents in Pharmacy Teaching & Learning

http://www.pharmacyteaching.com

Review article

Practical significance: Moving beyond statistical significance

Michael J. Peeters, PharmD, MEd, FCCP, BCPS*

College of Pharmacy and Pharmaceutical Sciences, University of Toledo, Toledo, OH

Abstract

Practical significance is an important concept that moves beyond statistical significance and *p* values. While effect sizes are not synonymous with practical significance, it is a basis for evidence of substantive significance. Investigators should find and report effect sizes whenever possible. To build evidence for practical significance in pharmacy education, three methods are discussed. First, effect sizes can be compared to *general interpretation guidelines* for practical significance. Second, using the effect sizes, investigators can *benchmark* by comparing effect sizes to external information from other studies; however, this information is not always available. Where prior data is limited, a third method after determining effect size is for investigators to calculate in their cohort an instrument's minimally important difference; the effect size could be compared to this *minimally important difference*, as opposed to a general interpretation guideline. A method to calculate the minimally important difference is described, as well as applications. Regardless, effect sizes must be determined and should be reported in articles; its comparator may vary as evidence for practical significance—so interpretation is key. Reporting effect sizes can enable benchmarking by others in the future and facilitate summaries through meta-analysis. It is clear that reporting evidence of practical significance with effect sizes is needed; simply reporting statistical significance is not enough. After reading this article, readers should be able to explain practical significance, recognize evidence of practical significance in other reports, and carry out their own analysis of practical significance using one or more of the methods described herein.

Keywords: Significance; Statistical significance; Practical significance; Effect size; Pharmacy education; Learning assessment

Situation

Finding practical significance is essential for investigating meaningful educational interventions and is often a more stringent criterion than statistical significance. One example of the importance of this distinction comes from a recent project in which faculty at our institution investigated the development of critical thinking among our PharmD students. Development of critical thinking is a broadly accepted goal of higher education,[1,2] including health professions education.[3,4] To measure development of critical thinking of PharmD students in our College of Pharmacy, we introduced periodic (approximately annual) longitudinal assessments using standardized critical thinking tests.

The director of educational assessment of the college asked that, as opposed to prior critical thinking studies, we move beyond evaluating statistically significant differences with these assessment findings. (*Note*: This director of educational assessment is already familiar with specific results and comparisons provided with standardized tests such as the Pharmacy Curriculum Outcomes Assessment and American Association of Colleges of Pharmacy surveys for PharmD programs.) Do these critical thinking tests show meaningful, practical measurement differences? If so, this College of Pharmacy may better evaluate whether the current curriculum is meaningfully helping to foster students' critical thinking development, and if not, where might curricular revisions be targeted?

* Correspondence to: Michael J. Peeters, PharmD, MEd, FCCP, BCPS, College of Pharmacy and Pharmaceutical Sciences, University of Toledo College of Pharmacy & Pharmaceutical Sciences, 3000 Arlington Ave, Mail Stop 1013, Toledo, OH 43614.

E-mail: michael.peeters@utoledo.edu

http://dx.doi.org/10.1016/j.cptl.2015.09.001

1877-1297/

Aside from this scenario, another example of the importance of distinguishing between statistical and practical significance comes from looking at a locally created annual assessment (i.e., progress examinations) of pharmacy knowledge among first-year, second-year, and third-year PharmD students and assessing whether students' knowledge development is fostered by a curriculum. The distinction might also be important if an educator sought to investigate a teaching and learning method such as a flipped classroom and evaluate students' learning using the course's final examination and comparing these results to those of last year's class.

Methodological literature review

Some statistically significant comparisons are too small for practical significance[5]; that is, some statistically significant results can be too inconsequential for meaningful, practical impact. Statistical significance is related to sample size, and p values are a common index used. Investigations with very large samples can detect very small differences, statistically, in an outcome between groups.[5] Another method to coax a statistically significant difference is to use an outcome measure with a large scale; larger scales can more easily show differences—but differences may not be meaningful either. The medical literature has numerous examples of statistically significant results that have questionable clinical significance, such as the use of topical diclofenac to improve pain control for osteoarthritis of the knee.[6] In this study, a quality-of-life instrument showed a statistically significant improvement over placebo for patients using topical diclofenac, though the change was so small numerically that other investigators have questioned its practical significance.[6]

Practical significance is contrasted with statistical significance. Practical significance in different contexts can be synonymous with substantive or clinical significance.[5,7–9] Unlike statistical significance that is simply a determinant of an inferential statistical test and has a formulaic process for interpretation using null-hypothesis significance testing (NHST), practical significance is less certain, and no single formulaic approach will always be best.[5] Teaching PharmD students about practical significance using a simple formulaic process is problematic as well. Evidence of practical significance can be sought using a few strategies. All use effect sizes. One strategy is to use effect sizes along with general interpretation guidelines. A second strategy uses effect sizes and benchmarking. A third strategy uses effect sizes and minimal important difference.

Using Method 1: General interpretation guidelines

Simply using effect sizes can suggest practical significance.[5–9] For example, Cohen's d is one popular effect size coefficient for quantifying a difference in education and other social sciences literature. It can be used to compare the means and standard deviations of two independent groups or two paired assessments from one cohort (e.g., pre- versus post-testing).[7] Cohen provided a general framework for interpreting these Cohen's d effect sizes [small (0.2), medium (0.5), and large (0.8)] and, similarly, for Pearson correlations [small (0.1), medium (0.3), and large (0.5)]; other significance test interpretations were provided as well.[10]

Using Method 2: Benchmarking

This is another strategy for determining practical significance. General interpretation guidelines have limits due to context. Thus, benchmarking uses comparative data from similar sources to set standards (i.e., expectations) based on the cumulative experiences of others. Benchmarking can provide real-world comparison instead of relying on more generic or distribution-based effect size interpretations, or even just using statistical significance. To this end, Hill et al.[11] provide an example of benchmarks for comparison within an educational setting, albeit for kindergarten through 12th grade and not focused on higher education; as such, this benchmarking reference may provide limited help. Another more rigorous example of benchmarking in pharmacy education (and briefly mentioned in this article's initial "situation") is the Pharmacy Curriculum Outcomes Assessment (PCOA), which fosters broad comparison with other colleges/schools of pharmacy. The PCOA is not without limitations either, as pharmacy students may not have fully completed all coursework included on the PCOA when they sit for this examination. The PCOA can also be limited by the broad contexts pooled from many different institutions, and so a smaller group of similar peer institutions may facilitate improved comparisons.

Using Method 3: Minimal important difference

In situations where further evidence of practical significance is desired beyond Method 1, and in the absence of literature or better context specifics with Method 2, a third method may be used. It uses standard error of measurement (SEM)[a] and is a statistical, distribution-based method to determine the *minimal important difference*.[12] Of note from the health measurement literature, use of *SEM* has been associated with clinical significance[12–14]; for health measurement instruments, the concept of statistical versus clinical significance has been investigated, and practical significance has been termed the *minimal clinically important difference*. For a multitude of health measurement instruments that will all use different scales, such as the

[a]This SEM should not be confused with "SE" (standard error) reported within many statistical programs; SEM is *different* from standard error of the mean (even though each can use SEM as their acronym).[15] Standard deviation, the standard error of the mean, the standard error of the estimate, and the standard error of the measurement are all conceptually distinct though related issues.

SF-36 for health-status, the COPD-specific St. George's Respiratory Questionnaire, or WOMAC for osteoarthritis pain/stiffness/physical dysfunction,[6] the SEM can be computed from the data set and can provide a different number for each individual scale; however, all have a defined ratio of 0.5 × SEM.[12–14,16,17] That is, once error because of the measurement instrument has been taken into account, practical difference will go beyond statistical significance.

Advantageously, SEM corrects for psychometric unreliability of educational testing and provides a level of confidence for decisions based on test results, regardless of scale length or sample size. Within a normal distribution, about 68% of data is ±1 standard deviation of the mean, and about 95% of data is ±2 standard deviations; likewise, ±1 standard error of measurement (SEM) would be a 68% confidence interval around the mean score, while ±2SEM would be a 95% confidence interval. Any scores within that ±1 SEM interval, while numerically different, are not outside of measurement error and should be regarded as *functionally the same*; to statistically discriminate students' ability within an assessment, scores >1 SEM would be meaningfully different from one another.

Situation recommendation

For the critical thinking situation described in the introduction, evidence for practical significance was sought. For each test, the first and second scores were quantified with a Cohen's *d* effect size. First, these Cohen's *d* values were compared to general interpretation guidelines. Second, limited prior data among pharmacy students provided a backdrop for analyses but was not overly helpful because of the small number of studies.[18] Third, a SEM-based cut-off for practical significance was used.[19] The formula for SEM is[15]

$SEM = SD \times SQRT$ [1−reliability], where SD is standard deviation and reliability is internal consistency.

For standardized tests that did not report reliability to users, an internal consistency of 0.75 was used, assuming conservatively that a standardized test should be at least acceptably reliable. The resulting SEM was one-half standard deviation (0.5 × SD), which corresponds well to prior literature of minimally important difference.[6,12–14,16,17] This one-half standard deviation is also a "moderate" effect size using Cohen's *d* (medium = 0.5).[10] *Thus (and for the director of educational assessment), using a Cohen's d effect sizes of "moderate" or greater should be the specific threshold for suggesting practical significance from these critical thinking assessments.* Differences with Cohen's *d* of less than 0.5 were deemed practically insignificant, although some scores could numerically increase to a small extent and show statistical significance. For the college's test results, differences in scores over time were to be compared using Cohen's *d* and interpreted as practically significant or not using SEM. Specific results regarding these critical thinking assessments have been reported elsewhere.[19]

Applications

Importantly, SEM use is suggested for educational testing.[15,20] Finding important, practical, educational significance could also use an SEM-based parameter approach as applied to the health measurement literature. Naturally, a plethora of locally-produced educational tests are employed in colleges/schools of pharmacy around the country, with their varied ranges of test content and test length derived from local program needs. This SEM-based method could be used to calculate a comparison index for each different test at each college/school, when other comparison strategies are insufficient or lacking.

The SEM also has application for high-stakes testing in any college/school of pharmacy. Using SEM to form confidence intervals around a cut-score provides a quantitative rationale to determine which scores (less than the cut-score) are functionally the "same" as the cut-score. For example, if the cut-score for a "pass" is 75% while the SEM is 5%, then any score greater than 70% (75–5%) should be a considered a "pass." Hays et al.[21] illustrate this use and its practical significance further.

For periodic assessments investigating student growth, an increase <0.5 times the standard deviation may not be impactful. A smaller effect size of 0.15 by Cohen's *d* may be only a statistical artifact from that investigation; it may not be replicated in future cohorts or at a different institution. In general guidelines, a Cohen's *d* of 0.15 would be considered a "trivial" effect size.[10] As noted above, we used this approach with critical thinking development.[19]

Similar to raters' judgments having few, coarse categories for ratings,[22] SEM illustrates confidence in score precision. Score results that are reported with precise decimal-place numbers and/or with multiple significant digits are likely to be *too* precise; this may give some readers false confidence in the precision of the reported results.[23] Limiting significant digits to two or using only whole rounded numbers may also encourage this etiquette with numbers.[24]

Using a Cohen's *d* cut-off of 0.5 for practical significance may seem stringent; one would omit some statistically significant findings. Fortunately, successful interventions in education have relatively large effect sizes. A meta-analysis of many published educational, psychological and behavioral treatments found a pooled mean Cohen's *d* effect size of 0.47[25]; that is, approximately half of the interventions were beyond one-half of a standard deviation, with some effect sizes well beyond 0.5. Also, educational effect sizes often appear larger when compared to clinical effect sizes.[26] Successful educational interventions will often have a medium or large effect size.[27] Within an investigation, a statistically significant finding may result from a data set's score differences only because of imprecision from the test instrument's measurement. In fact, if there is a small effect that investigators are convinced is real, then further research to better identify and describe it

would be much more worthwhile to the academic community than a simple, "See—it works!" Declaring that an intervention "works" can be of interest initially, but the natural next question is, "How well and to what extent does it work?" That is where the effect sizes come in and why confidence intervals are being promoted so much.

Implication

Described as "the new statistics"—effect sizes, confidence intervals and power calculations—these are not new to research although they may be novel to an individual investigator. These "new" statistics appear to be preferred to the traditional approach using statistical significance testing through NHST.[28] There is a considerable amount of literature among medical sources, psychology, and other social sciences on this subject of effect sizes and NHST.[27–36] While some authors have advocated for an outright change from the NHST to "the new statistics,"[28–30] others have suggested to simply include both *p* values and effect sizes.[7,8,31–35]

While effect sizes have been described in a few overlapping ways in the literature, they are conceptually simple and represent a magnitude [size] of a difference in the measured outcome (effect).[30] To aid readers' analyses and interpretations, investigators should report an effect size within every study.[5–8,15] Importantly, in their recommendations for the scholarly work in medical journals, the International Committee of Medical Journal Editors directs to "*avoid* relying solely on statistical hypothesis testing, such as *p* values, which fail to convey important information about effect size and precision of estimates" and "in particular, distinguish between clinical and statistical significance."[37]

Adding complexity, there are different categories of effect sizes: unstandardized and standardized. Unstandardized effect sizes are easiest to report and may not need further calculation beyond descriptive statistic reporting; an example is reporting the sample's difference between means (with standard deviations) for change, while other effect sizes are noted in the Table. Some authors have advocated unstandardized effect sizes as the preferred format to report effect sizes,[35] and this has been the most advocated approach in the medical literature as well.[33,36,38] With specific critiques such as different study designs, and correction for reliability or range restriction, those authors have illustrated that generic interpretations of standardized effect sizes have limitations due to context[35]; they should be interpreted with caution or, in those authors' opinions, avoided in lieu of unstandardized effect sizes. Meanwhile, the social sciences (education, psychology, political science, economics, and business/marketing, etc.) have taken a different approach and use mainly standardized measures. Within educational testing, advantages of standardized measures (especially the standardized mean difference of Cohen's *d*) include easier interpretation, regardless of a test's underlying test length and scoring scale. Standardized effect sizes, from potentially diverse instruments that may have used different test lengths and scoring scales, can be combined much more easily into a meta-analytic summary. As an example, a meta-analysis investigated instrument-based improvement in depression, where different instruments had been previously used in the literature. Investigators combined those effects to provide one pooled, summative effect size estimate.[39] As evidence-based medicine has suffused medical care, meta-analysis has become a powerful tool to integrate the evidence from multiple sources into quantitative summaries; it extends evidence-

Table
Some common effect size measures

Effect size measure	Description
Odds ratio	Quantifying how strongly an exposure is associated with treatment or risk; this is slightly different than risk ratio.
Risk ratio	Probability of event in exposed group compared to unexposed group; this is slightly different than odds ratio.
Risk reduction (absolute or relative)	Change in clinical risk or treatment in relation to control group—either absolute numbers or in a relative ratio of one another.
Number need to treat/harm	Inverse of absolute risk reduction; number of patients that need to be exposed to a treatment for one patient to benefit or be harmed.
Hazard ratio	From survival analysis, similar to relative risk; similar to relative risk ratio taking into account loss-of-sample by death.
Difference in group means	Absolute difference between groups (standard deviations needed to show distribution); most simple *unstandardized* measure.
r	Size of correlation; how closely associated two variables are.
R^2	Proportion of variance explained by a linear regression model.
Cohen's *d*	*Standardized* comparison between two group means, or between two paired means at different points in time.
η^2	Analogous to *r*-squared but for analysis of variance (ANOVA); *standardized* proportion of variance explained.

based medicine's conventional "gold standard" of the randomized clinical trial.[40,41] Thus, effect sizes, and not *p* values, are a basis from which evidence of substantive significance is developed[8] and from which data can be combined within meta-analyses.[42]

Limitations

Practical significance is *not* synonymous with effect sizes. Reporting a large effect size does not simply indicate practical significance; however, reporting an effect size for an investigation's finding is one way to provide

- Journal articles for this paper's references had been selected to highlight papers related to those specific concepts discussed.

A further primer on standard error of measurement

- Harvill LM. An NCME [National Council on Measurement in Education] Instructional Module on Standard Error of Measurement. *Educ Meas Issues Pract.* 1991;10(2):33-41.

Books

- Streiner DL, Norman GR. *Health Measurement Scales.* 4th ed, New York, NY: Oxford University Press; 2008.
    - While technical at times, this book (431pp) is one of the few psychometric-oriented in the health professions. Standard error of measurement and minimal clinically important difference are discussed.
- Grissom RJ, Kim JJ. *Effect Sizes for Research: univariate and multivariate applications*. 2nd ed. New York, NY: Routledge; 2012.
    - A comprehensive book (434pp) on the plethora of effect sizes for different research designs and purposes. These effect size discussions go well beyond the common Cohen's d or number needed to treat.
- Ellis PD. *The Essential Guide to Effect Sizes: statistical power, meta-analysis, and the interpretation of research results*. New York, NY: Cambridge University Press; 2010.
    - This book (173pp) broadens an effect size discussion to include confidence intervals, statistical power calculations, and meta-analysis.
- Cummings G. *Understanding The New Statistics: effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge; 2012.
    - This book (519pp) is focused on the "new statistics". The author has written much elsewhere on reporting these statistics instead of traditional null-hypothesis significance testing.

Box. Recommended resources.

evidence for concluding a practical significance.[31] Similar to validity, where multiple sources of validity evidence help demonstrate a conclusion of validity,[15] effect sizes provide evidence towards substantive, practical significance.

As discussed above, there are advantages and disadvantages to reporting either unstandardized or standardized effect size measures. While unstandardized measures are easier to calculate (in fact, many investigators will already have collected these before reporting), they are still within the units of their instrument's measurement, and the scales from multiple instruments cannot be combined easily for meta-analysis without conversion to standardized effect sizes. For standardized effect size indices, which can be combined across different instrument scales, this calculation takes further consideration.[b] *Regardless of which type of measure is used, it is clear that effect size should be reported along with any significance testing within research reports.*[5–8,15,28–37]

That said, contextual factors in educational settings could present challenges wherein benchmarking can become increasingly difficult. Establishing peer groups in an attempt to control for context may be warranted for better comparisons. In addition, comparisons of effect size across dissimilar studies should be done with caution. Of note, effect sizes do not necessarily indicate validity with an instrument; a new instrument may elicit a large effect size while measuring the wrong dimension of students' knowledge or ability. More is needed for valid conclusions such as validity evidence for content.[15]

Conclusion

Practical significance is an important concept that moves beyond statistical significance and p values. Practical significance most often uses effect sizes and is therefore not restricted by sample size. The threshold for practical significance cannot be overcome or improved by simply increasing the number of study participants, while statistical significance can be. Benchmarking can also provide evidence for practical significance, though the educational test will need to be the same (i.e., standardized) for most meaningful comparisons. By finding the minimal important difference, SEM can help determine practical significance and has shown prior utility with educational testing. Results greater than the SEM provide evidence for practical significance, while those less than the SEM may not be meaningfully significant. The inherent diversity of locally produced tests, with varied content and test length, can use this method. Reporting effect sizes and identification of practical differences is vital to advancing educational research. For additional related discussion, the Box provides further resources (Box).

[b] At the time of writing, free Cohen's d calculators are available online.

Author contribution

M.J.P. conceived, drafted, and revised this paper. Both Dr. Kimberly Schmude and anonymous peer-reviewers improved its communication. M.J.P. accepts responsibility for the paper's entirety.

Acknowledgments

The author thanks Dr. Kimberly Schmude for her helpful reviews in drafting and revising this manuscript.

Conflicts of interest

None to report.

References

1. Roksa J, Arum R. The state of undergraduate learning. *Change*. 2011;43(2):35–38.
2. Pithers RT, Soden R. Critical thinking in education: a review. *Educ Res*. 2000;42(3):237–249.
3. Ross D, Loeffler K, Schipper S, Vandermeer B, Allan GM. Do scores on three commonly used measures of critical thinking correlate with academic success of health profession trainees? A systematic review and meta-analysis. *Acad Med*. 2013;88(5):724–734.
4. Peeters MJ. Cognitive development of learners in pharmacy education. *Curr Pharm Teach Learn*. 2011;3(3):224–229.
5. Kirk RE. Practical significance: a concept whose time has come. *Educ Psychol Meas*. 1996;56(5):746–759.
6. Biacus C, Caraiola S. Effect measure for quantitative endpoints: statistical versus clinical significance, or "how large the scale is?". *Eur J Intern Med*. 2009;20(5):e124–e125.
7. Hojat M, Xu G. A visitor's guide to effect sizes. *Adv Health Sci Educ Theory Pract*. 2004;9(3):241–249.
8. Sullivan GM, Feinn R. Using effect size—or why the p-value is not enough. *J Grad Med Educ*. 2012;4(3):279–282.
9. Thompson B. "Statistical," "practical," and "clinical": how many kinds of significance do counselors need to consider? *J Couns Dev*. 2002;80(1):64–71.
10. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–159.
11. Hill CJ, Bloom HS, Black AR, Lipsey MW. Empirical benchmarks for interpreting effect sizes in research. *Child Dev Perspect*. 2008;2(3):172–177.
12. Norman GR, Sridhar FG, Guyatt GH, Walter SD. Relation of distribution-and anchor-based approaches in interpretation of changes in health-related quality of life. *Med Care*. 2001;39(10):1039–1047.
13. Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? *J Biopharm Stat*. 2004;14(1):97–110.
14. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol*. 1999;52(9):861–873.
15. Peeters MJ, Beltyukova SA, Martin BA. Improving rigor in scholarship of teaching & learning: a primer on educational testing and validity of conclusions. *Am J Pharm Educ*. 2013;77(9): Article 186.

16. Norman GR, Sloan JA, Wyrwich KW. Interpretation of change in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. 2003;41(5): 582–592.
17. Norman GR, Sloan JA, Wyrwich KW. The truly remarkable universality of half a standard deviation: confirmation through another look. *Expert Rev Pharmacoecon Outcomes Res*. 2004;4 (5):581–585.
18. Reale MC, Witt BA, Riche DM, Baker WL, Peeters MJ. Development of critical thinking among health professions students: a meta-analysis of longitudinal studies. *Am J Pharm Educ*. 2015;79(5): Article S4 [abstract].
19. Peeters MJ. Development in critical thinking: one institution's experience. *Am J Pharm Educ*. 2015;79(5): Article S4. [abstract].
20. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. *BMC Med Educ*. 2010;10:Article 40.
21. Hays R, Gupta TS, Veitch J. The practical value of the standard error of measurement in borderline pass/fail decisions. *Med Edu*. 2008;42(8):810–815.
22. Peeters MJ. Measuring rater judgment within learning assessments, part 1: why the number of categories matters in a rating scale. *Curr Pharm Teach Learn*. 2015;79(5):656–661.
23. Peeters MJ, Schmude KA, Steinmiller CL. Inter-rater reliability and false confidence in precision: using standard error of measurement within PharmD admissions essay rubric development. *Curr Pharm Teach Learn*. 2014;6(2):298–303.
24. Norman G. Data dredging, salami-slicing, and other successful strategies to ensure rejection: twelve tips on how to not get your paper published. *Adv Health Sci Educ Theory Pract*. 2014;19(1):1–5.
25. Lipsey MW, Wilson DB. The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *Am Psychol*. 1993;48(12):1181–1209.
26. Norman G. Editorial—the effectiveness and the effects of effect sizes. *Adv Health Sci Educ Theory Pract*. 2003;8(3): 183–187.
27. Sullivan GM. Is there a role for spin doctors in Med Ed research? *J Grad Med Educ*. 2014;6(3):405–407.
28. Cummings G. The new statistics: why and how. *Psychol Sci*. 2014;25(1):7–29.
29. Hubbard R, Lindsay RM. Why *p*-values are not a useful measure of evidence in statistical significance testing. *Theory Psychol*. 2008;18(1):69–88.
30. Kelley K, Preacher KJ. On effect size. *Psychol Methods*. 2012;17(2):137–152.
31. Fritz A, Scherndl T, Kuhberger A. A comprehensive review of reporting practices in psychology journals: are effect sizes really enough? *Theory Psychol*. 2012;23(1):98–122.
32. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. *Front Psychol*. 2013;4:Article 863.
33. Faraone SV. Interpreting estimates of treatment effects: implications for managed care. *Pharm Ther*. 2008;33(12):700–711.
34. Mays MZ, Melnyk BM. A call for the reporting of effect sizes in research reports to enhance critical appraisal and evidence-based practice. *Worldviews Evid Based Nurs*. 2009;6(3):125–129.
35. Baguley T. Standardized or simple effect size: what should be reported. *Br J Psychol*. 2009;100(Pt 3):603–617.
36. Fidler F, Thompson N, Cummings G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. *Psychol Sci*. 2004;15(2):119–126.
37. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. *International Committee of Medical Journal Editors Website*. ⟨www.icmje.org⟩ Updated December 2014. Accessed November 24, 2015.
38. Greenland S. Meta-analysis. In: Rothman K, Greenland S, eds. *Modern Epidemiology*. Philadelphia, PA: Lippincott-Raven; 1998:287–318.
39. Brown HE, Pearson N, Braithwaite RE, Brown WJ. Biddle SJH. Physical activity interventions and depression in children and adolescents. *Sports Med*. 2013;43(3):195–206.
40. Haynes RB. Of studies, syntheses, synopses, and systems: the 4S evolution of services for finding current best evidence. *ACP J Club*. 2001;134(2):A11–A13.
41. Berlin JA, Golub RM. Meta-analysis as evidence: building a better pyramid. *J Am Med Assoc*. 2014;312(6):603–605.
42. Murad MH, Montori VM, Ioannidis JPA, et al. How to read a systematic review and meta-analysis and apply the results to patient care. *J Am Med Assoc*. 2014;312(2):171–179.