Meta-Analytic Methodology and Conclusions about the Efficacy of Formative Assessment

Derek C. Briggs
Maria Araceli Ruiz-Primo
Erin Furtak
Lorrie Shepard
Yue Yin

August 20, 2012

pre-print, Educational Measurement: Issues and Practice

Abstract

In a recent article published in EM:IP, Kingston & Nash (2011) report on the results of a meta-analysis on the efficacy of formative assessment. They conclude that the average effect of formative assessment on student achievement is about 0.20 SD units. This would seem to dispel the myth that effects between 0.40 and 0.70 can be attributed to formative assessment. They also find that there is considerable variability in effect sizes across studies, and that only the content area in which the treatment is situated explains a significant proportion of study variability. However, there are issues in the meta-analytic methodology employed by the authors that make their findings somewhat equivocal. This commentary focuses on four methodological concerns about the Kingston & Nash meta-analysis: (1) the approach taken to select studies for inclusion, (2) the application of study inclusion criteria, (3) the extent to which the effect sizes being combined are biased, and (4) the relationship between effect size magnitude and characteristics of outcome measures. After examining these issues in the context of the Kingston & Nash review, it appears that considerable uncertainty remains about the effect that formative assessment practices have on student achievement.

Introduction

In the article, "Formative Assessment: A Meta-Analysis and a Call for Research," Kingston and Nash ([KN], 2011) present the results from a meta-analysis of studies that evaluate the effects of formative assessment on student achievement. After deriving 42 unique effect-size estimates from 13 studies that met their inclusion criteria, they compute a median effect size of 0.25. Using a random effects modeling approach, the authors subsequently estimate an overall mean effect size of 0.20. These results imply that the efficacy of formative assessment practices is much smaller than the frequently cited range of 0.40 to 0.70 that is often attributed to Black and Wiliam (1998a).[1] Furthermore, KN argue that the small number of experimental and quasi-experimental studies in the formative assessment literature makes it difficult to come to a more nuanced understanding of the factors that make some formative assessment practices more effective than others.

Many of the conclusions from the KN study are quite similar in nature to those provided in publications and presentations by Bennett (2011), Ruiz-Primo, Li, Yin, Morozov (2010), and Shepard (2005; 2009). It can be argued that what makes the KN study unique is the attempt to replace the "urban legend" of a 0.40 to 0.70 effect size with quantitative estimates from a rigorous and methodologically defensible meta-analysis. Yet while KN's results are intuitively plausible and their recommendations—such as calling for more experimental evaluations of formative assessment—are entirely commendable, their meta-analysis involved some questionable decisions. The purpose of this commentary is to discuss these decisions so that subsequent research in this area can more formally test the robustness of their findings.

---

[1] In their full review published in the journal *Assessment in Education*, Black and Wiliam (1998a) said specifically that they could not conduct a meta-analysis because the underlying differences among studies were so great as to make "amalgamations of their results" meaningless (p. 53). They pointed specifically to differences in assumptions about learning and to lack of attention to relevant variables that might explain variation among studies. Black and Wiliam did, however, cite numerous meta-analyses on narrower more focused topics, and they signaled their enthusiasm for large effects by picking .7 to illustrate, in their conclusions, the impressive educational benefits that such an effect size could imply. The oft repeated .4 to .7 range of effects comes from the more popularized summary of their review published in *Phi Delta Kappan* in the same year (Black & Wiliam, 1998b). These values rely heavily on a 1986 article by Fuchs and Fuchs, which focused on special education treatments. For more on this, see Bennett, 2010, p. 10-13.

In what follows we raise concerns about four methodological issues that could threaten the validity of KN's conclusions:

1. the approach taken to select studies for inclusion,
2. the application of inclusion criteria,
3. the extent to which the effect sizes being combined are biased, and
4. the relationship between effect size magnitude and characteristics of outcome measures.

The first two of these issues have to do with decisions that could change the estimate of the overall effect of formative assessment on student achievement; the second two have to do with decisions that would help to explain why the effect sizes vary across studies.

Study Retrieval and Inclusion Criteria

Searching for Studies

One of the most important but underappreciated aspects of conducting a meta-analysis is the iterative process initially required to cast a net as wide and as inclusive as possible to retrieve candidate studies. When a series of keywords are used to search for studies, how does one know that the optimal keywords are being used? Since educational researchers commonly use different terms to describe the same constructs or use the same term to describe different constructs[2], it is usually sensible for meta-analysts to use multiple synonyms when a literature search is being conducted. For example, in a meta-analysis of the effects of instructional innovations on STEM achievement in college settings, Ruiz-Primo and colleagues (Ruiz-Primo, Briggs, Iverson, Talbot, & Shepard, 2011) chose many different synonyms as keywords for the construct of interest and then validated the keywords in two ways. First, an initial list of keywords was developed and submitted for scrutiny to an advisory panel with domain-specific expertise. This procedure helped them to generate new keywords that had not been previously considered. Second, the research team generated a list of influential and highly regarded studies by directly contacting scholars recommended by the advisory panel. This list of

[2] This is a familiar problem in educational measurement that Kelley (1927) called the "jingle-jangle" fallacy, where two tests with the same name might measure different constructs, or two tests with different labels might measure essentially the same thing.

"word-of-mouth" studies was then used to test the adequacy of the search procedures. Searches of literature were conducted using combinations of the proposed keyword list. For each word-of-mouth study not retrieved, new keywords or search engines were added until the study in question could be retrieved. On these grounds, the authors were able to make a case for the validity of their study retrieval process.

Based on the information KN provide in their manuscript, it is difficult to make a similar case for the keywords used in their search. According to the authors, this process consisted of (1) retrieving the primary studies referenced by Black and Wiliam (1998a), (2) using the keywords "formative evaluation," "formative assessment," "formative test," and "assessment for learning" as search terms in conjunction with six databases (ERIC, PsychInfo, Proquest, Google Scholar, JSTOR, Dissertation Abstracts International), and (3) identifying and retrieving "relevant articles presented at conferences." First, this description does not provide sufficient information to fully replicate the authors' search, such as the total number of initial studies that resulted from all three sources, or the range of conference proceedings searched. Second, and more importantly, this description gives the impression that KN did not attempt to validate their choice of keywords, even though this could have been done using some subset of the studies referenced by Black and Wiliam (1998a).

Study Inclusion Criteria

After searching for candidate studies KN imposed filters using four criteria: use of a control or comparison group design, the study must take place in an academic K-12 setting, inclusion of appropriate statistics needed to calculate effect sizes, a publication date of 1988 or later, and an explicit description of the intervention as formative assessment or assessment for learning. Although the rationales for the first four of these criteria are largely self-evident, the rationale for the fifth is not. KN describe their rationale as follows:

> Black and Wiliam's (1998a) review of formative assessment literature included studies that involved several different learning theories, such as those that investigated the impacts of mastery learning approaches, curriculum-based measurement, as well as the effects of feedback, goal orientation, and self-

> assessment. While all of these methods certainly appear to fit, at least on some level, into the realm of formative assessment, the learning theories behind each method vary greatly. Due to these vast differences and our desire to quantitatively analyze results rather than summarize results, our focus was narrowed down to studies that explicitly use the word formative or the phrase assessment for learning to describe the process or assessments used. (p. 30)

Meta-analysis involves not only the computation of effect sizes to summarize results but also the examination of patterns of association between study characteristics and these effect sizes. The latter is what KN refer to as a "moderator analysis," something that they conducted by examining the extent to which variability in effect sizes could be explained as a function of three variables: content area, grade level and treatment type. One of the difficulties KN encountered was that because they had such a small sample of studies, conditioning on any single moderator variable often left a relatively small number of effect sizes to be compared. Yet a principal reason for the small number of studies may well be the restriction imposed by the fifth inclusion criterion. An alternative approach that could have resulted in a larger sample size would have been to view formative assessment as an umbrella term that is defined by the presence or absence of certain practices, such as teachers eliciting student thinking or providing informational feedback. Under this approach, heterogeneity in the learning theories that underlie formative assessment practices would not be viewed as a source of extraneous variability that needs to be minimized during study retrieval, but as variability one would wish to explain as part of a moderator analysis.

KN do not provide a convincing rationale for narrowing the domain of the formative assessment "construct" to the presence of a specific word or phrase. For example, should a study on the effect of providing feedback to students be excluded from review just because the treatment has not been given the explicit label of "formative assessment"? After all, "specific use of student feedback" represents one of the categories of KN's treatment type variable. This seems inconsistent. On the one hand, the authors are interested in whether studies with treatments that explicitly involve feedback are more effective than those that do not; on the other hand, they have imposed an inclusion restriction that could keep these studies out of the sample they wish to analyze.

Finally, one must keep in mind that while the starting date for KN's literature search was 1988, the phrases "formative assessment" and "assessment for learning" only became popular and widely used in the late 1990s and early 2000s (e.g., Black & Wiliams, 1998a; Chappuis & Stiggins, 2002; Stiggins, 2002). Hence this restriction would be expected to exclude relevant studies that were conducted between 1988 through the 1990s; in fact, there is only one study in KN's meta-analysis that was published before 2000.

Application of Study Inclusion Criteria

Once study inclusion criteria have been established, they must be applied in a consistent manner during the selection process. There is some evidence that KN did not apply their criteria consistently. Based on our own knowledge of the formative assessment literature we were immediately aware of two studies that were not included despite meeting KN's selection criteria. One study was conducted by Andrade et al. (2008) and another by Bonner (2009). The Andrade et al. study, which was published, in EM:IP, lists "formative assessment" as a keyword. While the Bonner study does not use formative assessment as a keyword ("formative evaluation" is used instead), this phrase does appear in the study's abstract.

Andrade et al.'s quasi-experimental study focused on the use of rubrics to facilitate students' ability to self-assess the quality of their essays. In both conditions students engaged in pre-writing essay activities (outlining, brainstorming), wrote a first draft, self-assessed their work, then revised and wrote a final draft of the essay. The key distinction in the treatment condition was the discussion and use of a rubric to guide a self-assessment after the first draft of the essay had been completed. The effect size for the Andrade et al. study (Hedges's g, the same approach used by KN; see next section) is 0.88.

Bonner's quasi-experimental study evaluated a summer professional development course in which teachers were trained to use practice tests to identify student learning weaknesses and plan appropriate feedback. The teachers involved in the study had already participated in summer "institutes" conducted to prepare high school students to

pass state examinations. The outcome measures used in the study were students' eventual scores on state tests in mathematics and biology. Students were exposed to the formative assessment treatment if they were tutored by a teacher who had been participating in professional development; students in the control group were tutored by teachers who did not receive professional development. When effect sizes for the Bonner study are computed using the same approach taken by KN (adjusting for the availability of pretest data using a standardized gain), the results are effect sizes of 0.63 in math and -.06 in biology.

While the Andrade et al and Bonner papers are examples of studies where inclusion criteria have not been applied consistently because they were incorrectly excluded, we also saw an example of a study that was incorrectly *included*. The study by Ruiz-Primo and Furtak (2007) was an exploratory, mixed-methods investigation of three teachers' everyday informal formative assessment practices. The authors used a four-step model of the formative assessment process as an analytic framework to aid in the interpretation of videotapes of three teachers leading classroom discussions (called "assessment conversations"). Teachers' informal formative assessment practices (as coded by the authors from classroom videotapes) were then linked to student achievement by means of a three-question outcome measure embedded in the teachers' regular curriculum. The comparisons being made in the Ruiz-Primo and Furtak study were used in a limited fashion to explore, in a purely correlational sense, possible links between informal assessment practices and student learning for the three teachers involved in their case study. Importantly, *there was no control group in this study*. Thus, inclusion of this study is inconsistent with the criterion established by KN that "treatment versus treatment comparison designs were not eligible, as an effect size derived from such as design would not represent the true impact of the treatment of interest without the use of pretest data" (p. 30).

Effect Size Computations and Bias

Significant variability exists in the effect sizes that were computed for each study. One possible explanation for a portion of this variability is the different ways that KN

computed effect sizes. The authors note that the effect sizes being computed depended upon the availability of pretest data. In the absence of pretest data, the effect sizes were computed using Hedges's g: "the difference between the treatment and control group means on the posttest outcome variable divided by the pooled standard deviation" (KN, p. 32). When pretest data were available, effect sizes were computed using "group differences at pretest and posttest divided by the pooled standard deviation." (KN, p. 32) (a standardized gain metric) [3]. From a methodological standpoint, neither of these statistics are necessarily unbiased estimates of the causal effect of a formative assessment intervention. Holding the underlying experimental design constant, an effect size that adjusts for differences in pretest scores among treatment and control groups will be preferable to one that does not. When these different effect size estimates are mixed together to arrive at an overall estimate across studies, it can be easy to forget that the within-study estimates vary with respect to their internal validity.

The results shown in Table 1 illustrate this point using the results for the six teacher pairs from Yin's (2005) randomized experiment that was part of the KN review. Effect sizes in the first column were computed using Hedges's g, ignoring available pretest data. The second column gives adjusted effect sizes using the available pretest data. While the latter column represents the effect sizes that were actually reported by KN, the former gives us some sense of the counterfactual had the pretest data not been available (as was apparently the case for a number of the studies under review).

[3] The denominator of pretest to posttest gain effect size is being computed by KN after pooling four unique SDs: the pretest and posttest SDs for the control group, and the pretest and posttest SDs for the treatment group. It is worth noting, however, that there are many other denominators that could have been used to compute the effect size when pretest data are available, and it would have been helpful for the authors to provide the reader for a rationale for their choice in this context. In contrast to the approach taken by KN, Becker (1988) suggested using the unpooled pretest SDs for treatment and control groups. Morris (2007) argued for the use of a pooled SD across treatment and control groups based only on pretest scores. Another defensible approach would be to only use the posttest SD for the control group. The latter approach seems most preferable in an educational testing context. When students are tested on material to which they have had little to no exposure to in the past, the pretest SD will be lower than the posttest SD, and this can lead to inflated effect sizes when only the pretest SD is used, or when both are pooled together.

Table 1. ***Effect Sizes on MC Test Outcome in Yin (2008)***

Teacher Pair	Effect Size: Hedges's g	Effect Size: Standardized Gain Prestest & Postest Pooled SD
1	-0.94	-1.07
2	-0.93	-0.58
3	-0.34	-0.20
4	0.13	0.55
5	-0.21	0.28
6	0.41	0.30
MEAN	-0.32	-0.14
SD	0.55	0.62

Notes: In both cases the mean effect size has been computed as a weighted average using posttest sample size.

The results shown in Table 1 indicate that it is highly likely that the effect sizes computed for studies without pretest data in the KN meta-analysis are biased relative to studies with pretest data. For example, even though the Yin study involved a randomized controlled experiment, adjusting for pretest data (going from column 2 to column 3) results in significant shifts in effect size estimates. The effect size for teacher pair 4 goes from 0.13 to 0.55; for teacher pair 5 the effect size goes from -0.21 to 0.28. On average across all six teacher pairs, the effect size is -0.14 with the pretest adjustment but -0.32 without the pretest adjustment. The reason for the shift upward when adjusting for pretest differences was that students in the treatment groups in the Yin study tended to be lower achieving than those on the control group. For studies in which students in the treatment condition were higher achieving at the outset of the intervention, an adjustment for pretest differences will have the opposite impact—it will shift the treatment effect downwards.

We can be sure that KN would have preferred, ideally, to have pretest data for every study, and we do not fault them for using the Hedges's g effect size computation when pretest data were not available. Our concern is that the authors do not make the reader aware that this limitation can have a significant impact on effect size magnitudes. Indeed, the reader is given no sense for which or how many of the effect sizes were

computed using Hedges's g, and which or how many were computed using standardized gains. This issue could have been explored by using effect size computation method as a moderator variable and/or by examining the empirical differences for studies where pretest data were available by approximating the counterfactual scenario as we have illustrated here. (Note that in the example shown in Table 1 the standard deviation of effect sizes computed using Hedges's g was smaller than the standard deviation computed using the standardized gain approach.) We think it is important for meta-analysts to help readers appreciate that the act of combining effect estimates from primary studies does nothing to guarantee that the estimates are not biased by threats to internal validity. Indeed, adjusting effect size estimates using pretest data is no inoculation against other omitted variables that may confound a quasi-experimental comparison.

The Impact of Outcome Measures on Effect Size Variability

The choice of outcome measure in an evaluative study of an educational intervention can have inflationary effects when they are too narrowly tailored to either the treatment or curriculum. Studies with outcome measures very close to the curriculum have been shown to exhibit higher effect sizes than outcome measures that are more distal (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). In other cases, a formative assessment itself may be almost indistinguishable from its associated outcome measure. Called by another name, this is the old problem of teaching to the test and the confounding of practice effects with learning. Though Black and Wiliam (1998a) did not focus on the relationship between features of the outcome measure and magnitude of effect, a basis for this concern certainly appeared several times in their review. For example, in summarizing the review by Rosenshine, Meister, and Chapman (1996) of students being taught to generate their own questions, Black and Wiliam observed larger effects when experimenters developed their own comprehension test. They also cited Khalaf and Hanna (1992), who found that in 16 of 20 studies summative tests contained questions similar to those in regular classroom tests.

To the extent that it was available, it would have been beneficial for KN to have provided readers with information about the characteristics and quality of the outcome measures that were employed across studies[4]. This information might have helped to resolve some of the interesting findings from KN's moderator analysis. For example, KN found that content area was a significant predictor of effect size magnitude, with the average effect size in English Language Arts about twice as large as the average for Mathematics and three times as large as the average for Science. If formative assessment interventions in English Language Arts tend to involve more outcome measures that overlap with the intervention (as in the Andrade et al study), while those in the domains of mathematics and science do not, this might explain why the mean effect size is higher for English Language Arts interventions.

Conclusion

We agree with the general conclusions made by KN at the end of their review: the hype and marketing of formative assessment has greatly outstripped the empirical research base that should be used to guide its implementation. A number of researchers have recently punctured the mythology behind the 0.40 to 0.70 average effect attributed to Black and Wiliam, and it is good for the field to have an empirical study that contributes to this debunking. We also echo their call for further research.

Where we part ways with KN is in regard to the new baseline of 0.20 their study would appear to establish for the average efficacy of formative assessment interventions. We believe that the methodological problems we have identified call the accuracy of this estimate into question. The validity of the inclusion criteria that were used to filter candidate studies has not been well-established and it is not clear that these criteria were applied consistently. If the effect sizes from a different sample of studies were used it would not be surprising to find that the average effect of formative assessment interventions is significantly lower *or* higher than 0.20.

[4] Indeed, it appears that they did in fact code studies for this information, but never used whatever variable(s) resulted from this coding in their subsequent analyses. On p. 31 KN write "A detailed coding sheet was designed to facilitate the recording of information from the studies. The information was coded on several dimensions: (1) sample descriptors, (2) research design, (3) nature of treatment, (4) dependent measure descriptors, and (5) effect size data."

Meta-analysis can play an important role in pushing research forward by not only summarizing an overall effect, but by helping researchers develop hypotheses in regards to factors that predict why some formative assessment practices appear to be more effective than others. KN's moderator analysis is a step in the right direction, but it missed some opportunities to push the field forward through the coding and analysis of other critically important moderator variables. For example, do formative assessment practices that are well aligned to certain theories of learning lead to larger effects on achievement than others? Do studies that control for pretest differences have significantly higher or lower effect sizes than those that do not? Are large effects an artifact of outcome variables that do not sufficiently generalize to the target domain of learning? We would encourage those conducting research on the efficacy of formative assessment practices to consider these issues in the future.

References

Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-references self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice*, *27*(2), 3-13.

Becker, B.J. (1988). Synthesizing standardized mean-change measures. British Journal of Mathematical and Statistical Psychology, 41(25), 257-278.

Bennett, R. E. (2009, June). *Formative assessment. A critical review*. Presentation at the University of Maryland, College Park, MD.

Bennett, R. E. (2011). *Formative assessment. A critical review*. *Assessment in Education: Principles, Policy & Practices*, 18(1), 5-25.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practices*, *5*(17), 7-74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139-144, 146-148.

Bonner, S. M. (2009). Investigating teacher use of practice tests for formative purposes. *Journal of MultiDisciplinary Evaluation*, 6(12), 125-136.

Chappuis, S. & Stiggins, R. J. (2002). Classroom assessment for learning. *Educational Leadership*. *60*(1), 40-43

Fuchs, L. S., & Fuchs, D., (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*, 199-208.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: World.

Khalaf, A. S. S., & Hanna, G. S. (1992). The impact of classroom testing frequency on high-school students' achievement. *Contemporary Educational Psychology*, *17*, pp. 71-77.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28–37.

Morris, S.B. (2008). Estimating effect sizes from pretest-posttest-control group designs, *Organizational Research Methods*, 11(2), 364-386.

Roshenshine, B., Meister, C., Chapman, S. (1996). Teaching students to generate questions: a review of the intervention studies. *Review of Educational Research*, 66, pp. 181-221.

Ruiz-Primo, M. A., Briggs, D., Iverson, H., & Shepard, L. (2011). Impact of undergraduate science course innovations on learning. *Science, 331*, 1269-1270.

Ruiz-Primo, M. A., Li, M., Yin, Y., Morozov, A. (2010, March). *Identifying Effective Feedback Practices on Student Learning: A Literature Synthesis*. Paper presented at the Annual Meeting for the National Association of Research in Science Teaching. Philadelphia, PA.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching*, *44*(1), 57-84.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. & Klein, S. (2002). On the evaluation of systemic education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369-393.

Shepard. L. A. (2005, October). Formative assessment: Caveat emptor. ETS Invitational Conference. The Future of Assessment: Shaping Teaching and Learning. New York, NY.

Shepard. L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice*, *28*(3), 32-37.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan. 83*(10), 758-65

Yin, Y. (2005). *The influence of formative assessment on student motivation, achievement, and conceptual change*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Yin, Y., Ayala, C., Shavelson, R. J., Ruiz-Primo., M. A., Brandon, P., Furtak, E., Tomita, M., & Young, D., (2008). On the measurement and impact of formative

assessment on students' motivation, achievement, and conceptual change. *Applied Measurement in Education*, *21*(4), *335-359.*