Title: Valuing a more rigorous review of formative assessment's effectiveness

Authors and Affiliations: Helen Apthorp, Mary Klute, Tony Petrites, Jason Harlacher, and Marianne Reale, Marzano Research

Abstract Body

Background / Context (*Description of prior research and its intellectual context*)**:**
Prior reviews of evidence for the impact of formative assessment on student achievement suggest widely different estimates of formative assessment's effectiveness, ranging from 0.40 and 0.70 standard deviations in one review (Black & Wiliam, 1998a; 1998b) to 0.20 in a meta-analysis (Kingston & Nash, 2011; 2015). One reason for the widely different estimates may be the way in which formative assessment was defined. Authors of the first review defined formative assessment as the process of "frequent feedback that students receive about their learning" (Black & Wiliam, 1998a, p. 7). Authors of the meta-analysis acknowledged that formative assessment is an umbrella term referring to many forms of strategies, such as "student-reflection activities, detailed student feedback, assessment conversations, and curriculum embedded assessment" (Kingston & Nash, 2011, p. 29). To be inclusive, authors of the meta-analysis decided to let authors of the synthesized reports define formative assessment. The topic relevance/intervention inclusion criterion for the meta-analysis was as follows: authors "explicitly use the word *formative* or the phrase *assessment for learning* to describe the process or assessments used" in the intervention being studied (Kingston & Nash, 2011, p. 30). One criticism of this definition, however, is that it narrows "the domain of the formative assessment 'construct' to the presence of a specific word or phrase" (Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012, p. 14).

Disagreement or confusion about defining formative assessment is not limited to researchers. Some or many practitioners misunderstand what formative assessment means or do not know how to use formative assessment effectively. Members of the REL Central Formative Assessment Research Alliance (FARA), including elementary school principals, high school teachers, and district administrators have observed that teachers "check for learning at the end of learning rather than during learning," and report they do not have enough time to analyze data and provide student feedback during instruction.

To address both researcher and practitioner confusion about formative assessment and its effectiveness for classroom practice, this poster proposal for SREE 2016 will present a systematic approach to defining, identifying and reviewing formative assessment interventions used as part of a systematic review of their effectiveness. Presenters will discuss how they grappled with reporting results in ways that would resonate with teachers and administrators responsible for allocating time and funds to formative assessment interventions and professional development. Additionally, presenters will discuss the role of impact variation in enhancing the utility of findings for both practitioners and researchers.

Purpose / Objective / Research Question / Focus of Study (*Description of the focus of the research)***:**
The purpose of the study is to describe variability in the effectiveness of formative assessment for promoting student achievement by refining, updating and applying Black and Wiliam's (1998a) original typology of formative assessment interventions and including only those studies that meet rigorous evidence standards for supporting causal inferences.

Setting (*Description of the research location)***:**
The systematic evidence review was conducted at the Regional Educational Laboratory (REL) Central administered by Marzano Research. REL Central is one of 10 Regional Educational Laboratory contracts funded and operated by the U.S. Department of Education Institute of

Education Sciences (IES). The REL program is authorized under the Education Sciences Reform Act (ESRA) of 2002 to conduct applied research and development and disseminate scientifically valid research in a manner that supports education decision making by practitioners and policy makers. REL Central, in the fourth year of a five-year contract for 2012 to 2017, supports the decision making of practitioners and policy makers in state and local education agencies in a 7-state region, including Colorado, Kansas, Missouri, Nebraska, North Dakota, South Dakota, and Wyoming.

Population / Participants / Subjects *(Description of the participants in the study: who, how many, key features, or characteristics)***:**
The systematic evidence review focuses on the effectiveness of formative assessment for improving the academic achievement of students in grades 1-6, including both regular education students and students with learning disabilities, emotional disturbance, and/or intellectual disabilities.

Intervention / Program / Practice *(Description of the intervention, program, or practice, including details of administration and duration)***:**
Formative assessment is defined as a process that engages teachers and students during instruction in continuous, systematic evidence gathering to improve student learning (Black & Wiliam, 2009; Chappius, 2009; Greenstein, 2010; Heritage, 2010; Marzano, 2010; Moss & Brookhart, 2009; Popham, 2011). The present review focuses on formative assessment in which the cycle of gathering and interpreting evidence occurs within a short or medium period of time (minutes, days, or weeks). While assessment information from end-of-course, end-of-grade, or other summative testing can be used formatively at any time, the utility of the shorter cycle is in adjusting instruction, while the utility of the longer cycle is in adjusting curriculum (Brookhart, 2014; Perie, Marion, & Gong, 2009).

Research Design *(Description of the research design)***:**
A systematic evidence review was used to identify what is known and what is not yet established about the efficacy of different formative assessment practices for improving student academic achievement. A comprehensive search of research literature and consultations with formative assessment researchers was conducted to identify potentially relevant studies. Researchers screened studies for relevance, reviewed eligible studies, and assigned evidence ratings using an approach modeled after the What Works Clearinghouse (WWC) *Procedures and Standards Handbook, Version 3.0*, for reviewing comparative group designs. To describe variation in effectiveness, researchers applied a coding system to classify formative assessment interventions, outcome assessments, student samples, and study contexts, and characterize study findings.

Data Collection and Analysis *(Description of the methods for collecting and analyzing data)***:**
Over 160,000 studies were identified through a keyword literature search and initial screen. Three additional phases of screening were conducted, yielding 152 studies eligible for a rigorous review and rating of evidence (see figure B.1). Researchers reviewed studies and recorded information using survey software and a coding form based on an adaption and extension of the *Studies on Formative Assessment Rating* form developed and used by McMillan et al. (2013). This project's study coding form captured basic information (such as study ID number, sample, and location relevance), type of study design and other methodological and descriptive features (such as sample participants and nature of control group), intervention descriptors, outcome

descriptors, and findings). For each study, WWC-certified researchers used an approach modeled after What Works Clearinghouse *Procedures and Standards Handbook, Version 3.0,* procedures and evidence standards for reviewing comparative group designs to rate studies and complete the data tabs in the WWC Study Review Guide (U.S. Department of Education, 2014). [1]

Researchers assigned eligible studies to one of three ratings: meets standards with reservations, meets standards without reservations, and does not meet standards. *Meets without reservations* is the highest rating a study could receive. These studies were conducted in a way that supports causal inferences about the intervention. Readers of these studies, with a high degree of confidence, can infer that a formative assessment intervention caused the reported results. *Meets with reservations* is the middle rating. These studies were conducted in a way that does not support attribution of cause solely to formative assessment. Finally, studies rated *does not meet standards* were not conducted in a way that was rigorous enough to support the interpretation that a formative assessment intervention caused the reported results.

Interventions were classified according to a system based on Black and Wiliam's (1998a) original three broad categories: (1) teacher-directed formative assessment practices, (2) student-directed formative assessment practices, and (3) multi-component formative assessment systems (Black & Wiliam, 1998a). This review refines and updates the typology to identify and distinguish six mutually exclusive types based on who or what primarily leads or takes ownership of the process (see figure B.2).

Outcome assessments were classified into one of three levels of alignment with the intervention based on prior research relating intervention effects to type of outcome assessment, including: (1) broadly focused standardized tests, such as a reading achievement test; (2) narrowly focused standardized tests, targeting a subdomain, such as passage comprehension in the reading domain; and (3) specialized topic tests "developed specifically for an intervention (such as a reading comprehension measure developed by the researcher for text similar to that used in the intervention)" (Hill, Bloom, Black, & Lipsey, 2008, p. 176). Study samples were characterized by whether student participants were identified as regular education students or as students identified as having a disability and receiving special education services. Study contexts were characterized by grade level and by subject area (mathematics, reading, writing, science, technology and engineering, music and physical education).

Using only those studies that authors rated as *met standards with or without reservations*, researchers characterized study findings according to a benchmark effect size, statistical significance, and the direction of effect into one of five mutually exclusive categories: statistically significant positive, substantively important positive, indeterminate, statistically significant negative, substantively important negative (see Figure B.3). Variability in the effectiveness of formative assessment for promoting student achievement was described by relating characterization of study findings to type of intervention, student sample, and outcome assessment in a series of cross-tabulations within and across subject area contexts.

[1] **The** Study Review Guide **was developed by the U.S. Department of Education,** Institute of Education Sciences **through its What Works Clearinghouse project and was used by the authors with permission from the** Institute of Education Sciences**. Neither the Institute of Education Sciences nor its contractor administrators of the What Works Clearinghouse endorse the content herein.**

Findings / Results *(Description of the main findings with specific details)***:**
Although findings from the study are currently under review, the process of preparing for and responding to reviews of the report posed several challenges and opportunities to consider possible solutions regarding communicating the value of a more precise estimate of formative assessment's effectiveness to practitioners and policy-makers. Details about the different challenges and solutions will be shared with our SREE audience. Four such challenge/solution pairs are presented below as follows:

Presenting rigorous evidence review results in terms that have meaning and utility for practitioners and policy makers	
Challenges	**Solutions**
Explaining different study ratings • "meets standards without reservations" • "meets standards with reservations" and • "does not meet standards"	Discuss the consequences of basing education decision making on studies that do not meet standards for supporting causal inferences
Establishing the meaningfulness of different effect sizes	Draw comparisons among effect size estimates for different classroom interventions; relate effect sizes to benchmark effect sizes from policy relevant research; translate effect sizes into an improvement index
Keeping reports short and accessible	Provide report sections, subheadings and white space for readers to navigate; state the single point of each paragraph in a good topic sentence
Being transparent without too much detail	Engage multiple reviews and reviewers and prepare responsive drafts in multiple cycles; use appendices to describe methods

Conclusions (*Description of conclusions, recommendations, and limitations based on findings)***:**
In conclusion, the formative assessment evidence review provides an opportunity to examine evidence that ordinarily is interpreted as strongly supporting the view that formative assessment is effective for improving student achievement. By extending conceptual frameworks for categorizing types of formative assessments presented in prior reviews and adopting rigorous standards to assess whether studies support causal inferences, this evidence review allows practitioners and researchers to examine relationships between effectiveness and both malleable and non-malleable factors. The poster will pose and address challenges to communicating findings from the review in ways that are intended to encourage bi-directional pathways between decision-making and evidence on the effectiveness of formative assessment.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P. & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, October*, 139–148. http://eric.ed.gov/?id=EJ575146

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation and Accountability, 21*(1), 5–31. http://eric.ed.gov/?id=EJ829749

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice, 31*(4), 13–17. http://eric.ed.gov/?id=EJ988994

Brookhart, S. (2014). *The essence of formative assessment: Creating a common understanding of the formative assessment process (Part one of a three-part series)* (Archived webinar). Retrieved July 11, 2015 from https://www.relcentral.org/news-and-events/the-essence-of-formative-assessment-creating-a-common-understanding-of-the-formative-assessment-process/

Chappius, J. (2009). *Seven strategies of assessment for learning.* Portland, OR: Pearson Assessment Training Institute.

Greenstein, L. *What teachers really need to know about formative assessment.* Alexandria, VA: ASCD.

Heritage, M. (2010). Formative assessment: Making it happen in the classroom. Thousand Oaks, CA: Corwin Press.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177.

Kingston, N. & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. http://eric.ed.gov/?id=EJ951173

Kingston, N. & Nash, B. (2015). Erratum. *Educational Measurement: Issues and Practice, 34*: 55. doi: 10.1111/emip.12075.

Marzano, R. J. (2010). *Formative assessment & standards-based grading.* Bloomington, IN: Author.

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research & Evaluation, 18*(2), 1–15. http://eric.ed.gov/?id=EJ1005135

Moss, C. M., & Brookhart, S. M. (2009). *Advancing formative assessment in every classroom: A guide for instructional leaders.* Alexandria, VA: ASCD.

Pai, H-H., Sears, D., & Maeda, Y. (2015). Effects of small-group learning on transfer: A meta-analysis. *Educational Psychology Review, 27*, 79-102.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5–13. http://eric.ed.gov/?id=EJ853799

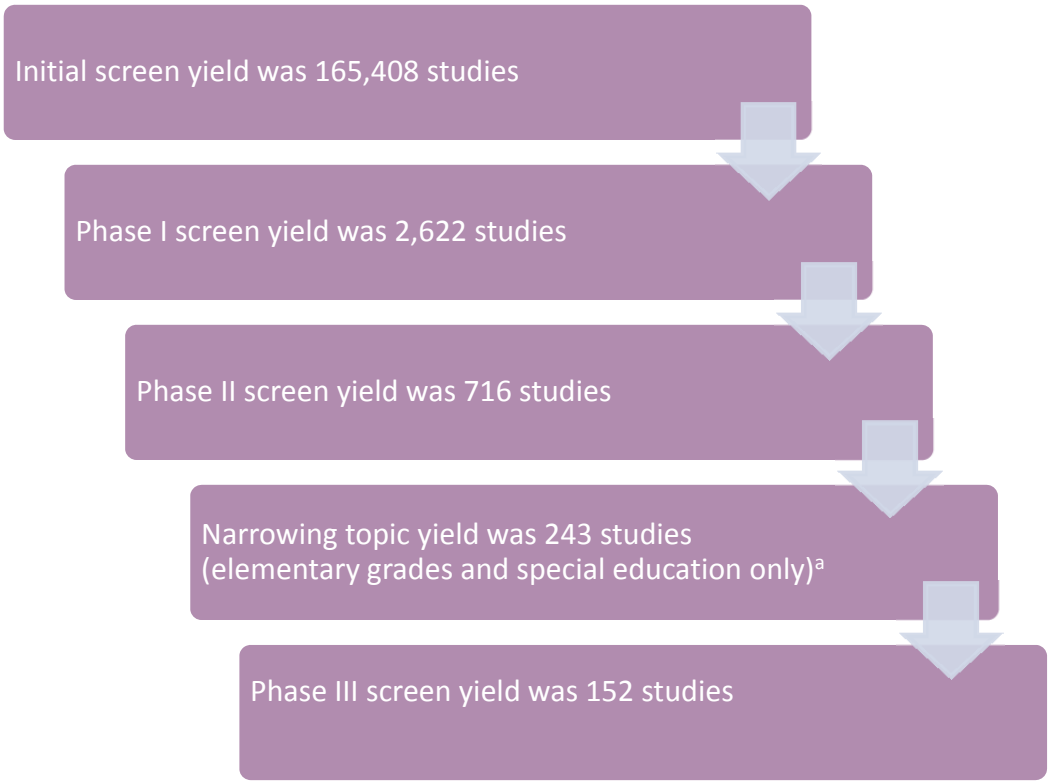
Popham, W. J. (2011). *Transformative assessment in action.* Alexandria, VA: ASCD.

Scammacca, N., K., Fall, A., & Roberts, G. (2015). Benchmarks for expected annual academic growth for students in the bottom quartile of the normative distribution. *Journal of Research on Educational Effectiveness, 8*(3), 366-379.

U.S. Department of Education (2014). What Works Clearinghouse™ *Procedures and Standards Handbook, Version 3.0.* Retrieved June 1, 2014 from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19

Appendix B. Tables and Figures

Figure B1. Formative assessment study yields from each phase of the study screening process



[a] Midway through Phase 2 screening, 473 studies with grade 7-12 students and/or only non-academic outcomes were excluded from further consideration. Some of these studies had already been reviewed and rated using What Works Clearinghouse *Procedures and Standards Handbook, Version 3.0*standards and some were in process, awaiting a second review or response to an author query.

Source: Author's compilation.

Figure B.2. Formative assessment intervention types

Student-directed. Students (1) appraise or monitor their own work or strategies or the work of their peers and (2) have the opportunity to reflect on the assessment information they gathered to determine next steps (Black & Wiliam, 2009).

Teacher-directed. Teachers (1) appraise or monitor student work, strategies, or progress and (2) have the opportunity to reflect on the assessment information they gather to determine next steps.

Computer-directed. Digital responses to questions or assignments are entered by a student or teacher and analyzed by programmed algorithms which generate an assessment score and/or other feedback contingent on the digital responses and stored data. This feedback is intended to inform next steps.

Teacher/student-directed. (1) Teachers appraise student work and students self-assess or peer-assess their work and (2) both teachers and students have opportunity to reflect on the assessment information to determine next steps.

Teacher/computer-directed. Students or teachers produce digital responses to software-generated questions or assignments, a software program analyzes the responses by programmed algorithms and generates an assessment score and/or other feedback contingent on the digital responses and stored data, and the teacher has an opportunity to inspect and reflect upon the assessment information to determine next steps.

Student/computer-directed. Students produce digital responses to software-generated questions or assignments, a software program analyzes the responses by programmed algorithms and generates an assessment score and/or other feedback contingent on the digital responses and stored data, and the student has an opportunity to inspect and reflect upon the assessment information to determine next steps.

Note: *Teacher* is a term used broadly to include any adult in charge of the teaching-learning event(s).

Figure B.3. Criteria for characterizing formative assessment effects that met WWC standards

Characterization	Criteria to meet rating, after applying any needed corrections
Statistically significant positive effect	• If one outcome, the estimated effect is positive and statistically significant • If more than one outcome in the same domain, the average effect size across the outcome measures is positive and statistically significant. • If more than one outcome in the same domain, and information to calculate effect sizes is not available, at least half of the effects are positive and statistically significant and no effects are negative and statistically significant.
Substantively important positive effect	• If one outcome, the estimated effect is greater than .25, but not statistically significant • If more than one outcome in the same domain, the average effect size across the outcome measures is greater than .25 but not statistically significant.
Indeterminate effect	• If one outcome, the estimated effect is between .25 and -.25 and not statistically significant • If more than one outcome in the same domain, the average effect size across the outcome measures is between .25 and -.25 and not statistically significant.
Substantively important negative effect	• If one outcome, the estimated effect is less than -.25, but not statistically significant • If more than one outcome in the same domain, the average effect size across the outcome measures is less than -.25 but not statistically significant.
Statistically significant negative effect	• If one outcome, the estimated effect is negative and statistically significant. • If more than one outcome in the same domain, the average effect size across the outcome measures is negative and statistically significant. • If more than one outcome in the same domain, and information to calculate effect sizes is not available, at least half of the effects are negative and statistically significant and no effects are positive and statistically significant.