View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by ScholarWorks@UMass Amherst

Practical Assessment, Research, and Evaluation

Volume 18 *Volume 18, 2013* Article 2

2013

Studies of the Effect of Formative Assessment on Student Achievement: So Much More is Needed

James H. McMillan

Jessica C. Venable

Divya Varier

Follow this and additional works at: https://scholarworks.umass.edu/pare

Recommended Citation

McMillan, James H.; Venable, Jessica C.; and Varier, Divya (2013) "Studies of the Effect of Formative Assessment on Student Achievement: So Much More is Needed," *Practical Assessment, Research, and Evaluation*: Vol. 18 , Article 2.
DOI: https://doi.org/10.7275/tmwm-7792
Available at: https://scholarworks.umass.edu/pare/vol18/iss1/2

This Article is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Practical Assessment, Research, and Evaluation by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

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 18, Number 2, February 2013 ISSN 1531-7714

Studies of the Effect of Formative Assessment on Student Achievement: So Much More is Needed

James H. McMillan, Jessica C. Venable & Divya Varier
Virginia Commonwealth University

Kingston and Nash (2011) recently presented a meta-analysis of studies showing that the effect of formative assessment on K-12 student achievement may not be as robust as widely believed. This investigation analyzes the methodology used in the Kingston and Nash meta-analysis and provides further analyses of the studies included in the study. These analyses suggest, consistent with other reviews, that some of the conclusions reached by Kingston and Nash may not be credible. The studies used in the Kingston and Nash meta-analysis were limited by the nature of the selection process, the questionable quality of their methodologies, and the multiple ways formative assessment was defined and operationalized, often without inclusion of recognized formative assessment characteristics that are needed for successful practice. These limitations mitigate Kingston and Nash's conclusion that the median effect size of experimental studies reviewed suggests a much smaller overall impact than reported by others. Recommendations for further research in this area are summarized to establish an improved body of literature on the effects of formative assessment on student achievement.

A recent series of articles has addressed the degree to which formative assessment affects achievement (Bennett, 2011; Filsecker & Kerres, 2012; Kingston & Nash, 2011). The recent meta-analysis by Kingston and Nash (KN), which investigated the relationship between formative assessment practices and student achievement, has drawn particular attention to its conclusion that the median effect size of this relationship is closer to .20 rather than the .40 - .70 range often cited. However, in light of Bennett's (2011) critical review of how formative assessment is operationalized and studied, published just 11 months earlier, a rejoinder to KN (Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012), and consideration of the Filsecker and Kerres (2012) analysis of formative assessment definitions and nature of evidence supporting its effect on achievement, it is appropriate to revisit KN's methodology and conclusions.

This article addresses some concerns and considerations that we believe will advance the understanding of the effect of formative assessment on student achievement by complementing the rejoinder provided by Briggs et al.. This is accomplished by examining in greater detail the methodological soundness of the studies used in the KN meta-analysis and analyzing the nature of interventions. More specifically, we maintain that weaknesses in the selection of studies for the meta-analysis, methodology, and variations in how formative assessment was operationalized in the studies, moderate the KN conclusions.

KN completed a meta-analysis of 13 studies that contained 42 effect sizes indicating the impact of formative assessment on student achievement. Their conclusion, not unlike Dunn and Mulvenon (2009), was that the overall quality of the research was not high ("A call for more high-quality studies is issued" [p. 28]). They also argued that both the weighted mean and median effect sizes "were significantly lower than previous estimates and customarily considered small

effect sizes," (p. 33). As pointed out by Briggs et al., these conclusions are consistent with others who maintain that the "hype and marketing of formative assessment has greatly outstripped the empirical research base that should be used to guide its implementation (p. 16). However, Briggs et al. also suggested that the methodology used by KN, as well as the nature of the outcome measures included, mitigate the conclusion.

This investigation addresses KN's findings in three ways: (a) in the selection criteria for the studies they included; (b) in the quality of the studies that were included; and (c) in the nature of the formative assessment practices that were included in the studies. Our intent is to provide a more systematic analysis of both the quality of the research that was included in the meta-analysis and the nature of the interventions. We hope this further analysis of the research will advance our understanding of the important causal relationship between formative assessment and student achievement.

Selection of Studies

KN used five criteria to determine inclusion of a study in their meta-analysis. The first was that the authors of the study "explicitly state that the intervention was formative in nature …or used the phrase 'assessment for learning,'" (p. 30). Hence, as long as "the authors stated that their intervention was formative …it was included in the analyses," (p. 30). This leads to the inclusion of a wide range of what many different researchers consider "formative assessment," and seems to us to make comparisons with previous research or with one another problematic. As we will show in our analyses, only a few of the studies employed interventions containing the generally accepted characteristics of effective formative assessment. These components include moment to moment interactions, communicating criteria for success, gathering information, providing feedback, and providing instructional adjustments or correctives (Filsecker & Kerres, 2009; McMillan, 2010). KN devised their own approach for operationalizing the nature of the formative assessment intervention, resulting in the overwhelming majority of studies being characterized as "Professional Development" (55%) or "Specific Use of Student Feedback" (17%) (p. 31). Bennett (2011) cautioned against making comparisons between studies that are, in fact, disparate. Briggs et al. also make a case that KN did not use a convincing rationale for their selection of terms to identify studies for the meta-analysis, citing examples of two studies that should have been included but were not.

Their second inclusion criterion, that participants were K-12th grade students, is appropriate, though given the range of grade levels with the relatively small number of studies included in each makes syntheses difficult. There may be legitimate differences in the nature of effective formative assessment at different grade levels, and KN addressed this issue. But treating formative assessment the same at different grade levels may not be reasonable.

The third inclusion factor focused on the nature of the design, excluding single group designs, and, apparently, alternative formative assessment intervention designs. One could argue that some single group pretest-posttest studies could be credible. More importantly, alternative treatment designs, which could compare different characteristics of formative assessment, could be informative.

The fourth criterion – that studies must include at least one quantitative measure of student achievement as a dependent variable -- is obviously critical. It would be interesting to know how many studies were excluded due to a lack of appropriate data from the authors. At the same time, KN included studies that measured student achievement as well as other variables, like student motivation (e.g., Yin, 2005). This may have affected the efficacy of the interventions, but KN made no attempt to distinguish the range of intent beyond investigating the relationship between formative assessment and student achievement. This issue is related to the final selection criteria: the use of a publication date of 1988 or later. It is surprising that only 13 studies since that time were identified. Black and Wiliam (1998) identified 250 studies, of a possible 681 publications, published after 1988, in their well-known review of research. It would be helpful to know more about the total number of studies first identified by KN in the search process, then the number of studies eliminated at each stage of the process and a

justification for the reasons cited here. Briggs et al. also suggest that the number of studies included was too small.

Methodological Quality of Studies

KN readily admitted that the studies they included could be of higher quality, though they do not elaborate on the "severe design flaws that would systematically bias the results" (p. 31), nor on which studies had design flaws. Cooper (2010) suggested that the quality of a meta-analysis is directly based on the quality of the studies included. According to Cooper, it makes little sense to include studies that may have significant methodological flaws. KN did not exclude or correct for what may have been low quality methodological designs. We contend that, in fact, most of the studies included in this meta-analysis were problematic from a methodological point of view, and that it doesn't make sense to pool results from poorly designed studies together and hope for a reasonable conclusion. This criticism was not addressed in the Briggs et al. rejoinder.

Formative Assessment Components Included in the Interventions

Because of the wide range of formative assessment components included in the studies it is difficult to understand how they should be grouped together in a meta-analysis. Although KN provide a very helpful analysis of different characteristics of the studies (e.g., grade level, content, type of intervention), there is no analysis of what we believe is critical to the effect formative assessment would be expected to have on student achievement – the depth and completeness of the intervention. Studies that include extensive formative assessment simply should not be expected to give the same result as an intervention that is fleeting or incomplete. In our analysis we have shown more specifically the extent to which critical components of effective formative assessment were included, or not, in the studies used for the meta-analysis.

A Further Analysis of the Studies

To investigate further both the methodological quality of the studies and the specific nature of the formative assessment components included in the interventions, we developed a set of guidelines to use in evaluating each of the studies (see Appendix 1). An initial draft of the guidelines was distributed to two outside experts in formative assessment for their review and suggestions. Following an initial trial evaluation of three articles, the guide was revised. Two reviewers examined each article. Following tabulations of the responses on the guideline, agreement between the reviewers was determined to identify specific methodological issues and the nature of the formative assessment components included in intervention. Based on the most salient themes in the studies, we generated the table found in Appendix 2, which highlights details of the studies that would add to the information already provided by KN. During our discussions, we found that, in several cases, insufficient information was provided about the study to allow us to classify the intervention based on some formative assessment components, such as "classroom culture" or "process attributes". We eliminated these categories in the table. In order to effectively present our findings, we collapsed components in the list into one column or category in the table. For instance, feedback related components were combined into one "Feedback Characteristics" column in the table, and findings on the duration and timing of interventions are summarized in a single column.

Quality of Methodology. To judge the methodological quality of the studies we used criteria suggested by the What Works Clearinghouse (2011) and threats to internal validity suggested by Valentine and Cooper (2008), and McMillan (2007). Each study was analyzed according to ten internal validity categories. Methodological concerns reflect those that, in our judgment, were serious enough to suggest that there existed a plausible threat to internal validity. In each case where such a threat is present, the results may essentially be uninterpretable due to one or more serious flaws in the methodology.

Formative Assessment Components. To evaluate the extent to which components of formative assessment were included in each study we relied on definitions provided by Bell and Cowie (2001), Bennett (2011), McMillan (2010), Ruiz-Primo and Furtak (2007), and Wiliam, Lee, Harrison, and Black (2004). Our initial list of components was then reviewed by

outside experts in formative assessment, and revised given their input. The components that were analyzed are included in Appendix 1. Our goal was to generate salient characteristics that would address the depth, complexity and duration of commonly accepted formative assessment practices as implemented in the studies under review. As such, the components list served as a guide to gain a deeper understanding of the intervention characteristics that were included by KN.

Findings

The results of our analyses are summarized in Appendix 2. It is immediately evident, from our judgments, that serious methodological concerns are present in most of the studies. Overall, most of the studies relied on quasi-experimental designs, with little other than the use of pretest scores to address the threat to selection.

The most common methodological problems were the unit of analysis, selection and instrumentation. KN explained that they included studies employing non-randomly assigned groups "out of necessity" (p. 30). Even with our liberal benchmark in making methodological judgments -- if the groups' pretest scores were similar we did not indicate selection as a threat to internal validity – five of the 13 studies still had notable selection flaws. Selection was judged to be a serious threat for five studies even with pretesting (Koedinger, McLaughlin & Heffernan, 2010; Poggio, Poggio & Glasnapp, 2007; Rackoczy, Klieme, Burgermesiter & Harks, 2008; Wiliam, Lee, Harrison, & Black, 2004; Yin, 2005), including all of the 21 classroom comparisons in the Wiliam et al. (2004) study. In at least five of the studies (Brookhart, Moss, & Long, 2008; King, 2003; Ruiz-Primo & Furtak, 2007; Touminen, 2008; Van Evra, 2003), students were nested in classrooms, although nesting was not accounted for in the analyses. The studies that used computerized assessment and feedback could also have had unique classroom-level effects. Additionally, we could not be assured that reliable and valid scores were used in three studies (King, 2003; Koedinger et al., 2010; Wiliam et al., 2004).

There is also a lack of emphasis on treatment fidelity in five studies (Boulet, Simard, & Demelo, 2012; Tomita, 2008; Touminen, 2008; Wiliam et al., 2004; Yin, 2005). We noted there was little to no careful documentation of systematic implementation, and we cannot be sure what formative assessment practices, intended or not, were actually used in the classrooms. This is particularly true for the effect sizes listed for teachers in the Wiliam et al. (2004) study. Considering that 21 of the 42 effect size indices are from this research report, the KN conclusions may be weighted too heavily by these particular effect sizes. The article did list different components of formative assessment reported by teachers in their action plans, including those contained in our definition of effective formative assessment. But the practices that were actually implemented, and to what extent, were not reported. This issue may have been compounded by the problematic control groups used for comparisons. These flaws were noted in spite of the fact that the magnitude of effects were among the strongest reported in the meta-analysis.

The variation in what was implemented, both contextually and in what formative assessment components were included, suggests to us that a meta-analysis of them may not be appropriate at all. Many of the studies (e.g., King, 2003; Rackozy et al., 2008; Touminen, 2008; Van Evra, 2003; Yin, 2005) examined the relationship between formative assessment and student beliefs, teacher beliefs and practices, motivation, and self-efficacy. Achievement, while a dependent variable, was not always the main focus of the study. This may have mitigated the effectiveness of these interventions in impacting achievement.

We also considered the inclusion of studies that had seemingly disparate purposes. For example, Touminen (2008) measured student achievement outcomes, though the primary focus of the study was how teachers collaborated with each other to reflect on and refine classroom practices. This study had two layers - teachers engaging in formative assessment practices for each other, and teachers engaging in formative assessment practices for their students. We questioned, then, why Touminen would be included in the meta-analysis with a study like Boulet et al. (2012), which was only interested in student achievement, or Van Evra (2003), that looked at student achievement and motivation but not teacher behavior or

perceptions. KN acknowledged that mean effect sizes "may not be meaningful if aggregated over studies with vastly different independent or dependent variables" (p. 29). The authors' solution was to conduct moderator analyses. Yet even with this procedure, we argue that the chosen moderator variable "Treatment Type" was insufficiently diverse to capture the full contexts of the studies.

We found that the reported level of detail about the formative assessment components implemented, generally, was insufficient. Much more specificity in describing the nature of the interventions would be helpful. The studies KN reviewed represented only a fraction of research on formative assessment. Yet even given the limited information available from the manuscripts, it is apparent from our analysis that the studies sampled stressed gathering data and providing feedback, with little emphasis on instructional correctives.

Just as instructional correctives are viewed as essential to formative assessment, there are many who consider student involvement to be important for effectiveness (McManus, 2008). Student involvement can be in the form of peer assessment or self-assessment. We found five studies from those included in the KN meta-analysis in which the formative assessment intervention included some form of student self-assessment (King, 2003, Ruiz-Primo & Furtak, 2007; Touminen, 2008; Van Evra, 2003; Wiliam et. al., 2004). In these investigations, feedback encouraged students to reflect on their understanding of what was being learned through conversation, debate, and revision, though self-assessment in the Wiliam et al. study was determined by the number of times a teacher mentioned self-assessment in the action plan for formative assessment. Only Tuominen's (2008) study involved the use of peer collaboration.

The role of feedback is critical in formative assessment (Filsecker & Kerres, 2012; McMillan, 2010). All 13 studies included some form of feedback, but Poggio et al. (2003) did not provide enough information for us to be able to characterize the feedback. However, there were several differences in the level of specificity in the nature of feedback provided to students. For instance, studies differed on the timing of the feedback: while some teachers provided immediate feedback to students (e.g., Koedinger et al., 2010; Yin, 2005), others provided delayed feedback (e.g., Boulet et al., 2012). Second, differences were noted among the studies in relation to whether feedback was delivered to individual students (Brookhart et al., 2008; Rich, Harrington, Kim, & West, 2008; Van Evra, 2003) or to a group (Rackoczy et al., 2008; Ruiz-Primo & Furtak, 2007; Tomita, 2008; Yin, 2004). Boulet et al. (2012) provided both group and individualized feedback.

Discussion

Although KN employed a statistically sophisticated meta-analysis to investigate the effect of formative assessment on K-12 student achievement, several weaknesses in their methodology, along with limitations in the quality of the studies, mitigates their conclusions. The selection of studies for a meta-analysis is critical, and it appears that several factors may have impacted the credibility of the selection process.

The most significant shortcomings of the KN meta-analysis, in our view, were: (a) their lack of attention to the studies' methodological quality; and (b) their lack of consideration of the specific nature of formative assessment under investigation in each study. Overall, only a few of the studies, in our opinion, had sufficient methodological rigor to justify inclusion in a meta-analysis. This is disheartening considering the presumably large number of studies eliminated from inclusion by KN in their meta-analysis. Furthermore, we were struck by how vague many of the studies were in their descriptions of the interventions. As a consequence, it is difficult to know whether the investigators rigorously touched on the hallmarks of what is generally accepted to be effective formative assessment. For example, there seemed to be little attention given to implementing instructional adjustments, a feature that is probably critical to improving student achievement. Gathering data about student understanding and providing feedback, without instructional adjustments, may not be a very powerful intervention. This is consistent with admonitions by Wiliam (2010) and Wiliam and Leahy (2007), among others, that instructional adjustments are essential to

effective formative assessment. Several of the studies emphasized feedback provided *after* instruction, rather than *during* instruction. As emphasized by Filsecker and Kerres (2012), this is an important distinction. When formative assessment is focused on progress of learning as students are instructed and process feedback, it is clearly most powerful. Heritage (2012), in agreement with Bell and Cowie (2001) and others, as summarized by Filsecker and Kerres, contend that the essence of formative assessment "is to enable teachers to respond to student learning in order to enhance that learning *while the student is in the process of learning*" (p. 182; emphasis added). Given that many of the studies in the KN meta-analysis did not contain this feature, it is not unreasonable to argue that the conclusions are dubious. Finally, there was also very little emphasis in the studies on student self-assessment and reflection.

Had there been more detail about the interventions, it is possible that KN would have changed their operational definitions of the four types of treatment emphases. We agree with Bennett (2011) and Young and Kim (2010) that definitional issues concerning formative assessment are problematic, notwithstanding Filsecker & Kerres (2012). At the very least, researchers need to clearly conceptualize and operationalize what formative assessment characteristics are used in their studies. That will be essential to being able to synthesize studies to gain a more holistic perspective about the effect of different components of formative assessment.

It is noteworthy that, according to our analysis, a single study in this meta-analysis (Rich et al., 2008) had no serious methodological flaws. Students received specific and intensive feedback over one academic year, though there was no indication of instructional adjustments. While we were not sure that the intervention included all components of effective formative assessment, the design was solid, and it showed effect sizes of +.30 and +.35. This study underscores two important points. First, the gains noted by Rich et al. should be considered important. It is a common misperception that Cohen's initial guidelines for interpretation of effect size (e.g., .2-.3 "small," .4-.6 "medium," and .8+ "large") are what should be used for evaluating the practical significance of educational interventions (Cohen, 1988). KN use Cohen's original guidelines, then provide a very helpful analysis of how effect size differences translate into improvements in the percentages of students rated proficient or above. If one could see a rise of 9-12% in this category, as they claim, most would interpret that to be very important. However, the main conclusion of their study – that both the mean and median observed effect sizes are "customarily considered small" (p. 33) -- rests on Cohen's guidelines. Whether from previous research or the present, it is important in our opinion to provide an appropriate interpretation of effect size.

Second, this study shows the difficulty in distinguishing patterns between methodological quality, formative assessment characteristics, and effect size. We attempted to find such patterns across all the studies, and were unable to detect any consistent trends. For example, larger effect sizes were not associated with formative assessment occurring multiple times over the entire academic year, nor were smaller effect sizes associated with short duration studies. We were also unable to ascertain if a type of internal validity threat, or lack thereof, was associated with effect size. This supports the contention that it is difficult to conduct this type of analysis and establish reasonable causal conclusions (Shepard, 2010).

Our reexamination of KN's methods and the quality and nature of the studies included suggests that any conclusions based on this meta-analysis about the relationship between formative assessment and achievement are tentative at best. We clearly agree with Briggs et al. that the suggestion that an effect size of .20 is closer to reality than higher effect sizes, is not credible. It also suggests that researcher efforts to investigate formative assessment need significant improvement. It is very difficult to advance our understanding of formative assessment with studies that lack credibility. Higher quality studies are needed. This is not an easy task given the applied nature of the research.

In summary, our field is advanced by KN's meta-analysis, but there is much work yet to be done. With clearer conceptual definitions, higher quality studies, and attention to all aspects of formative assessment, we will continue to build an important foundation about

the effect of formative assessment on student achievement.

References

Andrade, H., & Cizek, G. (Eds.) (2010). *Handbook of formative assessment.* New York: Routledge.

Bennett, E.B. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy and Practice. 18*(1), 5-25. doi: 10.1080/0969594X.2010.513678

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*, 536-553.

Beretvas, S. N. (2010). Meta-analysis. In G. R. Hancock & R. Mueller (Eds.) *Quantitative Methods in the Social and Behavioral Sciences: A Guide for Researchers and Reviewers* (pp. 255-263). New York: NY: Routledge.

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

Boulet, M.M., Simard, G., &Demelo, D. (1990). Formative evaluation effects on learning music. *Journal of Educational Research, 84*, 119 – 125.

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., & Shepard, L. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice, 31*(4), 13-17.

Brookhart, S. M., Moss, C. M., & Long, B. A. (2008). Teacher inquiry into formative assessment practices in remedial reading classrooms. *Assessment in Education: Principles, Policy & Practice, 17,* 41- 58.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Cooper, H.M. (2010). *Research synthesis and meta-analysis* (4th ed.). Thousand Oaks, CA: Sage.

Filsecker, M., & Kerres, M. (2012). Repositioning formative assessment from an educational assessment perspective: A response to Dunn & Mulvenon (2009). *Practical Assessment, Research & Evaluation, 17*(16). Retrieved from http://pareonline.net/getvn.asp?v=17&n=16

Heritage, M. (2012). Gathering evidence of student understanding. In J. H. McMillan (Ed.). *Sage Handbook of Research on Classroom Assessment* (pp. 179-196). Thousand Oaks, CA: Sage Publications.

King, M. (2003). *The effects of formative assessment on student self-regulation, motivational beliefs and achievement in elementary science.* Doctoral dissertation. Dissertations & Theses: Full text database (Publication No. AAT 3079342). Retrieved from http://proquest.umi.com/pqdlinkVer=1&Exp=10-22-2016&FMT=7&D1765252141&RQT=309&attempt=1&cfc=1

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

Koedinger, K.R., McLaughlin, E.A., & Heffernan, N.T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Educational Computing Research, 43* (4), 489 - 510.

McManus, S., Ed. (2008). *Attributes of effective formative assessment.* Washington, DC: Council of Chief State School Officers.

McMillan, J.H. (2010). The practical implications of educational aims and contexts for formative assessment. In H.L. Andrade & G.J. Cizek (Eds.), *Handbook of formative assessment* (pp. 41-58). New York: Routledge.

McMillan, J.H. (2007). Randomized field trials and internal validity: Not so fast my friend. *Practical Assessment Research & Evaluation. 12*(15). Retrieved from http://pareonline.net/pdf/v12n15.pdf

Poggio, A., Poggio, J., & Glasnapp, D. (2007, April). *The utility and impact of online computerized formative and early warning assessments on student performance.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Rakoczy, K., Klieme, E., Burgermeister, A., & Harks, B. (2008). The interplay between student evaluation and instruction. *Journal of Psychology. 216* (2), 111- 124.

Rich, C.S., Harrington, H., Kim, J., & West, B. (2008, March). *Automated essay scoring in state formative and summative writing assessment.* Paper presented at the meeting of the American Educational Research Association, New York, NY.

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*, 57-84.

Shepard, L. A. (2010). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice, 28*(3), 32-37.

Tomita, M.K. (2008). *Examining the influence of formative assessment on conceptual accumulation and conceptual change.* Doctoral dissertation. Dissertations & Theses: Full text

database (Publication No. AAT 3343949). Retrieved from http:// gradworks.umi.com/33/43/3343949.html

Tuominen, K.R. (2008). *Formative assessment and collaborative learning with support involving middle school mathematics teachers.* Doctoral dissertation. Dissertations & Theses: Full text database (Publication No. AAT 3302776). Available at gradworks.umi.com/33/02/3302776.html

Valentine, J.C., & Cooper, H.M. (2008). A systematic and transparent approach for assessing methodological quality of intervention effectiveness research: The study design and implementation assessment device (Study DIAD). *Psychological Methods, 13*(2), 130-149.

VanEvera, W.C. (2003). *Achievement and motivation in the middle school classrooms: The effects of formative assessment feedback.* Doctoral dissertation. Dissertations & Theses: Full text database (Publication No. AAT 3110082).

What Works Clearinghouse Procedures and Standards Handbook, Version 2. (2011). Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Wiliam, D. (2010). An integrative summary of the research literature and impicatios for a new theory of formative assessment. In H. A. Andrade & G. J. Cizek (Eds.), Handbook for formative assessment (pp. 18-40). New York: Routledge.

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.). *Formative assessment: Theory into practice* (pp. 29-42). New York: Teachers College Press.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on Student achievement. *Assessment in Education, 11* (1), 49 - 64.

Yin, Y. (2005). *The influence if formative assessments on student motivation, achievement, and conceptual change.* Doctoral dissertation. Retrieved from http://gradworks.umi.com/31/86/3186430.html

Young, V. & Kim D. (2010). Using assessments for instructional improvement: A review of the literature. *Education Policy Analysis Archives,* 18(19), 1-40. Retrieved from http://epaa.asu.edu/ojs/article/view/809

Appendix 1

Guidelines for Evaluating Studies used in Kingston and Nash (2011) Meta-Analysis

Studies on Formative Assessment Rating Form

4/9/12

Basic information
Date:
Coder:
Journal/Source:
First author:
Year: Volume: Issue:

STEP 1: Screening: ***(adapted from What Works Clearinghouse)***

	Yes	No
Does the study have an eligible design? (RCT/QE)	RCT QE	
Is the study a primary analysis of the effect of an intervention?		
Is the intervention professional development?		
Is the intervention a program, product, policy, practice that shows alignment with formative assessment?		
Does the study address at least one student achievement outcome? What is the outcome measure used? _____		
Comments:		

STEP 2: Assessing methodological quality of studies

Instruments	Not applicable	**Not controlled** does not meet evidence standards	**Partial control** may meet evidence standards	**Controlled** meets evidence standards
Reliability evidence		1	2	3
Validity evidence		1	2	3
Experimental Validity				
History		1	2	3
Experimenter effects		1	2	3
Intervention fidelity		1	2	3
Instrumentation		1	2	3
Selection (equivalence of groups)		1	2	3
Subject effects (e.g., please experimenter)		1	2	3
Confounding factors		1	2	3
Attrition		1	2	3
Unit of analysis		1	2	3
Diffusion of treatment		1	2	3
Other:		1	2	3
Fatal flaw(s) in design?	Yes	Maybe	No	
Comments:				

STEP 3: Formative assessment characteristics in the study:

		If yes, rate level of emphasis*		
Origin	Yes/ No/ Not known	Low	Medium	High
Primarily teacher-directed	Y N NK	1	2	3
Primarily student-initiated	Y N NK	1	2	3
combination	Y N NK	1	2	3
Timing				
Informal Formative Assessment • *Improvisational, spontaneous, arises out of teaching, information gathered is transient, flexible responses*	Y N NK			
Formal Formative Assessment • *Planned, precise data collection; focuses on specific aspect of learning*	Y N NK			
Combination of informal and formal formative assessment	Y N NK			
Takes place before instructional unit	Y N NK			
Takes place after instructional unit	Y N NK			
Takes place during instructional unit	Y N NK			
Takes place at a number of time points	Y N NK			
Process attributes				

Criteria for success: Learning goals/expectations are communicated to students (McManus, 2008).	Y N NK	1	2	3
Feedback: *specificity*				
Individualized (given to individual student)	Y N NK	1	2	3
Group (given to a group of students)	Y N NK	1	2	3
Check student comprehension (right/ wrong)	Y N NK	1	2	3
Elaborate on student understanding.	Y N NK	1	2	3
Offers suggestions on how student can improve.	Y N NK	1	2	3
Feedback linked to learning goals/expectations.	Y N NK	1	2	3
Feedback: *timing*				
Delayed.	Y N NK			
Immediate.	Y N NK			
Students given time to reflect on feedback before making changes/revisions.	Y N NK			
Feedback: *format*		Minimal		Extensive
Oral.	Y N NK	1	2	3
Written.	Y N NK	1	2	3
Student involvement		Low		High
Involves student self-assessment.	Y N NK	1	2	3
Involves peer-assessment.	Y N NK	1	2	3
Instructional adjustments/correctives				
Planned/prescriptive.	Y N NK	1	2	3
Unplanned/flexible.	Y N NK	1	2	3
Classroom culture	Y N NK	1	2	3
Teacher-student interactions are collaborative.	Y N NK	1	2	3
Teachers and students participate in choice of task.	Y N NK	1	2	3
Comments:				

*Low-level emphasis- very little, limited emphasis on the specific aspect of formative assessment
High-level emphasis- very extensive, heavy emphasis on the specific aspect of formative assessment

Appendix 2

Summary of Methodological Quality, Formative Assessment Characteristics, Outcome, and Effect Size.

Study	Methodology		Formative Assessment Characteristics					Outcome or Measure of Interest	Effect Size (Range) (*d*)
	Nature of Intervention	Design Quality Concerns: Plausible Threat to Internal Validity	Formal or Informal?	Feedback Characteristics	Instructional Correctives	Teacher and Student Involve-ment	Duration and Timing of Feedback		
Boulet, Simard, & Demelo (2012)	Teachers provided either oral, written or no feedback to secondary level music students.	**Confounding factors**: Written feedback was individualized, but oral feedback was given to group. It is hard to interpret if score increases were a result of individualized feedback or type of feedback. **Treatment fidelity**: No information on teachers, and how consistently they provided feedback to the students.	**Formal**: Type of feedback given was planned. Feedback type was based on pretest.	**Written** feedback: individualized and specific messages were conveyed to students. **Oral** feedback: given to the group and included providing correct answers to test items.	**Planned and prescriptive**: written feedback included a work plan for students who then worked independently; oral feedback included providing students with correct answers to the test items.	Teacher – directed.	One time feedback given on pretest performanc e in a previous course.	Student achieve-ment.	+.19
Brookhart, Moss, & Long (2008)	Primary teachers participating in extensive professional development instituted more systematic data collection, record-keeping, feedback, and data use over an academic year.	**Diffusion of treatment**: Control teacher practices could have been influenced by intervention teachers. **Unit of analysis**: Intervention implemented by classroom but students were used as unit of analysis.	Formal and informal	Extensive, specific, immediate and delayed feedback was provided to students.	Unclear. Scant, anecdotal indications that instructional correctives were made.	Teacher-directed.	Study takes place over the academic year; feedback given during instruction.	Student achieve-ment.	+.06, +.48
King (2003)	Teachers provided daily check-ups of student understanding with student reflection, including quizzes.	**Confounding factors**: Unknown factors possibly confounded with intervention and control classes **Diffusion of treatment**: Intervention and control classes in same school. **Instrumentation**: Student achievement measured using	Formal and informal Unknown	Daily check-ups of student understanding with student reflection and self-assessment and extensive, individualized feedback.	No indication that instructional correctives were made.	Teacher- and student-directed.	Study takes place over 4 weeks; 20 hours of intervention ; feedback given during instruction.	Student achieve-ment. Instruction. Classroom climate.	-.24

Study	Methodology		Formative Assessment Characteristics					Outcome or Measure of Interest	Effect Size (Range) (*d*)
	Nature of Intervention	Design Quality Concerns: Plausible Threat to Internal Validity	Formal or Informal?	Feedback Characteristics	Instructional Correctives	Teacher and Student Involve-ment	Duration and Timing of Feedback		
		several teacher-made tests. **Experimenter effects**: Researcher was a staff member at the school. **Unit of analysis**: Intervention implemented by classroom but students were used as unit of analysis.							
Koedinger, McLaughlin & Heffernan (2010)	Automated, online feedback provided with follow up questions and scaffolded lessons based on incorrect test question responses.	**Instrumentation**: Pre and post tests were different. **Selection**: Control students had higher pretest scores than intervention group.	Formal	Students given immediate, corrective feedback.	Unclear. Vague language indicated that teachers used data from the online system for correction.	Teacher- and student-directed.	Study takes place over academic 1 year for students responding to at least 60 items (two hours of content); immediate feedback provided during learning.	Student achieve-ment.	+.08
Poggio, Poggio & Glasnapp (2007)	Researchers analyzed Grades 5, 8 and 10 student data from a computerized formative assessment system and the state exams to examine the impact of formative assessment.	**Selection**: Since the study used and available large-scale dataset, there may have been reasons that determined whether and to what extent students were administered formative tests. Group equivalence was determined from demographic information, however, key factors like school emphasis on formative assessment, is unknown.	**Formal**: Assessments aligned with state tests and were administered to students via computer.	Unknown	Unknown	Unknown	Multiple assessments given at multiple time points.	Student achieve-ment.	+.09, +.19, +.25
Rackoczy, Klieme, Burgermesister & Harks (2008)	Researchers looked at how teacher evaluative and informational feedback influenced motivation and achievement	**Selection**: Groups of students not equivalent. Also it is unclear whether there was a comparison group.	**Informal**. There was no planned script in providing feedback.	Oral (?) Feedback was given in the classroom while students worked in groups and was based on work students engaged in	**Unplanned, flexible**: depending on the child's response, teacher behavior was recorded as corrective feedback and informational feedback.	Teacher-directed	Feedback provided in one classroom session during instruction.	Student achieve-ment. Motivation.	.00

Study	Methodology		Formative Assessment Characteristics					Outcome or Measure of Interest	Effect Size (Range) (*d*)
	Nature of Intervention	Design Quality Concerns: Plausible Threat to Internal Validity	Formal or Informal?	Feedback Characteristics	Instructional Correctives	Teacher and Student Involve-ment	Duration and Timing of Feedback		
	for secondary level students.			during the instruction period that was recorded.					
Rich, Harrington, Kim & West (2008)	Automated, online feedback based on writing samples.	None	Formal	Specific, extensive feedback provided to students to help them improve writing.	Unclear. Scant, anecdotal indications that instructional correctives were made.	Teacher-directed.	Study takes place over one academic year.	Student achieve-ment.	+.30, +.35
Ruiz-Primo & Furtak (2007)	Researchers observed differences in teachers' use of formative assessment strategies within the ERSU model on single science unit.	**Unit of analysis**: ESRU model tested in 3 classrooms but intervention effects are measured independently for each student in the classes.	**Informal**: Model being tested determined degree to which teachers elicited information, processed student understand-ing, and responded unrehearsed.	**Oral**: Model emphasized classroom conversations. Teachers looked for student understanding, used data and provided feedback to advance learning.	**Unplanned conversations**: Strategies employed varied, including promoting debate, referring to previous learning, and encouraging students to elaborate on thinking.	Interventio n model relied on **student-teacher interactio ns**. Feedback given to both individual students and class (whole-class discussions).	Implementa tion took place over 1 school year. Assessment conversatio ns were ongoing, not time-limited.	Student achieve-ment.	+.92
Tomita (2008)	Embedded assessments, which were reflective and encouraged students to think and debate through science concepts, were included in the FAST curriculum to see if they would result in higher student achievement as compared to a group that received only the FAST curriculum with no formative	**Experimenter effects:** Researcher was in same school as classrooms. **Diffusion of treatment**: The same teacher taught both control and experimental group. **Treatment fidelity**: Although only the experimental group engaged in reflection, control group students were not discouraged from discussion and explanation of concept.	**Formal**: Thorough planning of how reflective activity would incorporate formative components; teacher's role was well-defined well; observation s and	Feedback was primarily in the form of facilitating students to debate and reflect on science concepts.	**Unclear.** Study indicated that the Teacher facilitated discussion in a prescriptive manner based upon the FAST curriculum, however the nature of the correctives is unknown.	Teacher facilitated, but primarily student initiated.	Feedback was provided during reflection activity within instructional period. Instructiona l period is unknown.	Student achieve-ment.	+.09

Study	Methodology		Formative Assessment Characteristics					Outcome or Measure of Interest	Effect Size (Range) (*d*)
	Nature of Intervention	Design Quality Concerns: Plausible Threat to Internal Validity	Formal or Informal?	Feedback Characteristics	Instructional Correctives	Teacher and Student Involve-ment	Duration and Timing of Feedback		
	assessment.								
Touminen (2008)	Teacher-made formative assessment was supported by peer collaboration.	**Treatment fidelity**: No information was provided about formative assessment practices implemented in study classrooms. **Unit of analysis**: FACTS model was tested in 31 pre-Algebra sections, students were used as unit of analysis.	**Teacher/Peer formative assessments**: **formal.** Volunteer teachers observed classes and wrote common assessments, which were discussed and used to guide instructional adjustments. **Student formative assessments**: **unknown.**	**Oral and written** feedback was given to teachers. Teacher-to-teacher collaboration and the effects of those interactions on teaching practices were central to the study.	Teachers collaborated to suggest classroom improvements based on student mastery data.	Primarily teacher-directed. Some observations noted heavy student involvement.	Collaboration meetings took place at a number of time points. Assessment data collected at a number of time points.	Teacher beliefs about and practices of formative assessment. Student achieve-ment.	+.08
Van Evra (2003)	Study investigated the effects of receiving a range of written feedback on all classwork and homework assignment sin science versus just receiving completion scores.	**Experimenter effects**: Researcher was also the teacher implementing the intervention in both the experimental and control conditions. **Unit of analysis**: The study was implemented in four science classes, yet students were used as the unit of analysis.	**Formal**: Assessment was planned, and type of feedback was prescribed.	4 types of written feedback: performance, strategy, corrective, constructive. Given to students given based on individualized needs.	Provided corrective feedback that, for example, questioned the students' thinking and conclusions.	**Teacher-directed**. However, students could ask follow-up questions	Study took place over four weeks. Data collected at a number of timepoints.	Student achieve-ment. Motivation.	+.71
Wiliam, Lee, Harrison, & Black (2004)	Teachers received professional development on formative assessment and outlined features they would like to incorporate in their classrooms. Researchers observed teachers'	**Instrumentation**: Different, local tests were used to measure student achievement. **Selection**: Multiple haphazard comparison groups. **Treatment fidelity:** No information was provided about the formative assessment practices used by each teacher in the study.	Unknown. No details were provided on teachers' actual classroom practices under this study.	Teachers' action plans included questioning, feedback, sharing criteria with learners, and self assessment. Unknown which practices were actually implemented and to what extent.	Unknown. No details were provided on teachers' actual classroom practices under this study.	Unknown. No details were provided on teachers' actual classroom practices under this study.	No uniform implementation. Researchers noted that changes occurred towards the end of first school year; Unclear when feedback was provided.	Student achieve-ment. Teacher professiona lism.	-.35 - +1.55

Study	Methodology		Formative Assessment Characteristics					Outcome or Measure of Interest	Effect Size (Range) (*d*)
	Nature of Intervention	Design Quality Concerns: Plausible Threat to Internal Validity	Formal or Informal?	Feedback Characteristics	Instructional Correctives	Teacher and Student Involve-ment	Duration and Timing of Feedback		
	formative assessment practices and adaptation.								
Yin (2005)	Embedded assessments were included to guide students in their understanding of science concepts.	**Selection**: Author acknowledges there were issues with group equivalence even though groups were matched. **Treatment fidelity**: teachers took varied amount of time to complete the embedded assessments in both individual sessions (41-80 min) and total length (24-83 days).	**Formal**. Experimental group teachers received training on developmental trajectories and teaching strategies from researchers.	Feedback was immediate, given to a group. Feedback elaborated on student understanding.	**Planned and prescriptive**: facilitated classroom discussion based on formative, embedded assessments.	Teacher-directed	Assessments implemented at multiple time-points, but during specific instructional periods; feedback provided during instruction.	Motivation. Student achieve-ment. Conceptual change.	-1.07 - +.30

Citation:

McMillan, James H., Venable, Jessica C., & Varier, Divya (2013). Studies of the Effect of Formative Assessment on Student Achievement: So Much More is Needed. *Practical Assessment, Research & Evaluation*, 18(2). Available online: http://pareonline.net/getvn.asp?v=18&n=2

Corresponding Author:

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