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.

Volume 14, Number 7, March 2009 ISSN 1531-7714

A Critical Review of Research on Formative Assessment: The Limited Scientific Evidence of the Impact of Formative Assessment in Education

Karee E. Dunn & Sean W. Mulvenon
University of Arkansas

The existence of a plethora of empirical evidence documenting the improvement of educational outcomes through the use of formative assessment is conventional wisdom within education. In reality, a limited body of scientifically based empirical evidence exists to support that formative assessment directly contributes to positive educational outcomes. The use of formative assessments, or other diagnostic efforts within classrooms, provides information that should help facilitate improved pedagogical practices and instructional outcomes. However, a review of the formative assessment literature revealed that there is no agreed upon lexicon with regard to formative assessment and suspect methodological approaches in the efforts to demonstrate positive effects that could be attributed to formative assessments. Thus, the purpose of this article was two-fold. First, the authors set out to clarify the terminology related to formative assessment and its usage. Finally, the article provides a critical analysis of the seminal literature on formative assessment, beginning with Black and Wiliam (1998), and extending through current published materials.

The implementation of No Child Left Behind (NCLB) in 2002, and subsequent sanctions for lower performing school systems, has led to a myriad of educational interventions to improve student achievement. A common method advocated to improve student achievement is the use of formative assessments, both to improve the pedagogical practices of teachers and to provide specific instructional support for lower performing students. An almost unchallenged belief in education is that research has conclusively demonstrated that the use of formative assessment facilitates improvement in instructional practices, identifies "gaps" in the curriculum, and contributes to increased student performance.

However, as part of a series of studies being designed to evaluate the assessment and methodological practices used in "data driven decision-making," a review of the literature revealed limited empirical evidence demonstrating that the use of formative assessments in the classroom directly resulted in marked changes in educational outcomes. Basically, what began as a perfunctory review of literature on formative assessments for a manuscript on statistical methods, evolved into a critical analysis of both the operationalization of formative assessment and the methods employed to document the impact of formative assessments.

It is difficult to hypothesize, and somewhat irresponsible to conclude that the use of formative assessments does not provide information to help improve instructional practices or student outcomes in classrooms. This manuscript provides a critical examination of the formative assessment literature in particular issues related to the formative assessment lexicon, Black and Wiliam's (1998) seminal work, and more recent research. Finally, this manuscript provides the foundation for a series of manuscripts on "best practices" for evaluating student achievement through the use of formative assessment.

Review of the Literature

Over the past several years, a growing emphasis on the use of formative assessment has emerged, yet formative assessment has remained an enigma in the literature (Black & Wiliam, 1998; Leung & Mohan, 2004). When reading formative assessment literature and focusing on the issue of solidifying a definition of the term, an interesting and problematic theme arose. Formative assessment and its various manifestations (i.e. self-assessment, peer-assessment, and interim assessment) were defined not only by inherent characteristics, but also by the use of the assessment. Formative assessment's status as an ethereal construct has further been perpetuated in the literature due to the lack of an agreed upon definition. The vagueness of the constitutive and operational definitions directly contributes to the weaknesses found in the related research and dearth of empirical evidence identifying best practices related to formative assessment. Without a clear understanding of what is being studied, empirical evidence supporting formative evidence will more than likely remain in short supply.

For example, Black and Wiliam (1998) defined formative assessment as "all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (p. 10). Where as, the Council of Chief State School Officers (CCSSO) defined formative assessment differently according to the specifications provided by Formative Assessment for Students and Teachers (FAST), a department of CCSSO. FAST defined formative assessment as a process used during instruction to provide feedback for the adjustment of ongoing teaching and learning for the purposes of improving student achievement related to instructional objectives (Melmer, Burmaster, & James, 2008). In 2006, Popham stated that an assessment is formative to the degree that the information collected from the assessment is used during the assessed instruction period to improve instruction to meet the needs of the students assessed. In 2008, Popham defined formative assessment as a planned process during which the teacher or students use assessment-based evidence to adjust ongoing learning and instruction. Without any inter- or intra-individual consensus as to what the term formative assessment means, it is difficult to have a well-formed body of research.

To further complicate the issue of operationalizing formative assessment based upon the assessment itself as well as the use of evidence from the assessment, formative assessments serve a myriad of feedback related purposes such as diagnosis, prediction, and evaluation of teacher and student performance (Black & Wiliam, 1998). For example, Perie, Marion, and Gong (2007) argue that assessment issues can be clarified if assessment is defined by its purpose. From this perspective formative assessment is defined as assessment used by teachers and students to adjust teaching and learning, as compared to interim assessment that informs policymakers or educators at the classroom, school, or district level. Defining assessments in this fashion leaves a great deal of confusion for those trying to publish or consume assessment literature because one assessment could be used by students and teachers to inform the learning process as well as by administrators to create policy changes.

Moreover, a great deal of assessment literature is aimed at delineating between formative and summative assessment, yet summative assessment can be used for formative purposes (Bell & Cowie, 2000). It is important to note that we acknowledge that the purpose for which any assessment is developed and validated is an important aspect of assessment. However, a test that was designed to give formative feedback is only formative if the teacher uses it to provide feedback for the student. If the teacher only uses the formative assessment to provide a grade, is that assessment still formative? By most definitions the mere assessment of performance into a grade category (i.e., "A" or "B") is formative because it provides information on the achievement of the student and may be used for future instructional interventions. However, is this what is intended by the various definitions?

Although an assessment may be designed and packaged as a formative or summative assessment, it is the actual methodology, data analysis, and use of the results that determine whether an assessment is formative or summative. For example, Wininger (2005) used a summative assessment as a formative assessment by providing both quantitative and qualitative feedback about the results of the exam. Wininger (2005) called this formative summative assessment. This article exemplifies the complications that arise when one defines an assessment by its usage. An assessment is an assessment, and the manner in which an assessment is evaluated and used is a related but separate issue.

We do not argue that evaluation or use of assessment driven data is an unimportant aspect of the

assessment process. However, by operationalizing assessment as something unique from evaluation, researchers and educational stakeholders alike may begin to speak the same language related to the usage of these assessments and produce better research and more powerful academic outcomes. For example, a hammer is a hammer regardless of how it is used. If a hammer was defined by its use, it would make the discussion of the tool much more difficult in the remodeling of a home. It is easier to simply ask for and receive a hammer than to provide a description of the intended use (i.e., if you ask for an item that can make a hole in the wall, you might receive a sledge hammer in lieu of a hammer).

By separating the nature of an assessment from the use of its results, our perspective extends back to Scriven's (1967) original presentation of formative evaluation. Scriven (1967) described formative evaluation as the evaluation of an ongoing and malleable educational program. It was Bloom (1969) who attempted to transfer the term formative from evaluation to assessment. Perhaps this is where an understanding of the process of defining formative assessments first became complicated. The authors argue that defining formative assessment as a test and formative evaluation as the specific use of assessment data (be it formative or summative data) is more amenable to both classroom application and academic discourse.

Thus, the authors proposed that formative or summative assessment data may be evaluated and used for formative or summative purposes. The purpose of this manuscript was two-fold. The first aim was to provide a clear and more user-friendly terminology related to formative and summative assessment by reintroducing the concept of formative evaluation to the literature and separating it from assessment. Then, the authors briefly reviewed Black and Wiliam's (1998) seminal piece on formative assessment to dispel the myth that formative assessment had thorough empirical evidence supporting its positive impact on student performance followed by a review of more recent research.

Formative Evaluation as Opposed to Formative Assessment

In this section, the authors reintroduce and redefine formative evaluation as well as separate the issue of assessment from the issue of evaluation of assessment-based data. The authors define summative and formative assessment as well as summative and formative evaluation of assessment-based data. Summative assessments are those assessments designed to determine a students' academic development after a set unit of material (i.e., assessment of learning) (Stiggins, 2002). Formative assessments are assessments designed to monitor student progress during the learning process (i.e., assessment for learning) (Chappuis & Stiggins, 2002).

Although assessments may be designed for formative or summative purposes, the authors argue that resultant data may be interpreted either formatively or summatively. The authors further argue that the early mentioned definitions of formative and summative assessments that include how the data is used leads to issues in the literature due to the possibility of evaluating or using either type of assessment data formatively or summatively. The authors define the terms formative evaluation and summative evaluation in terms of the use of assessment data and separate the issue of assessment instruments from assessment use.

For our purposes, summative evaluation was defined as the evaluation of assessment based data for the purposes of assessing academic progress at the end of specified time period (i.e., a unit of material or an entire school year) for the purposes of establishing a student's academic standing relative to some established criterion. Formative evaluation was defined as the evaluation of assessment-based evidence for the purposes of providing feedback to and informing teachers, students, and educational stakeholders about the teaching and learning process. Formative evaluation also informs policy, which then affects future evaluation practices, teachers, and students. The reciprocal relationship between policy and formative assessment is graphically represented by the Key Model for Academic Success (See Figure 1). This model supports Shepard's (2000) assertion that it is not necessary to separate assessment from teaching; instead, teaching practices can and should be informed by and coincide with assessment practices and outcomes.

Having defined what is meant by "formative evaluation," it is important to separate the issue of assessment from the issue of formative or summative evaluation. In doing so, we hope to provide a more clearly defined nomenclature to frame the investigation of both effective implementation of formative evaluation as well as the effect of formative evaluation on student performance. Assessments are instruments for collecting information, in this case information

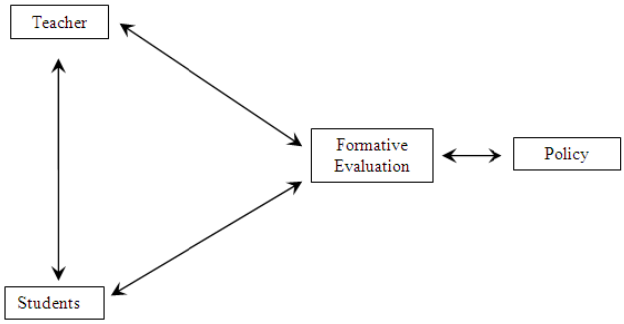

Figure 1. Key Model for Academic Success

about students' academic performance, including the actual learning process. Evaluation is a separate, but related issue that has to do with the use of assessment-based data. Although an assessment may be designed to be formative or summative, the data acquired by the administration of either type of assessment may be used for formative or summative purposes. In other words, formative evaluation or summative evaluation may be applied to either formative and summative assessment data. What arises from the application of this assessment lexicon is a practical model for a system of assessment and evaluation (See Figure 2).

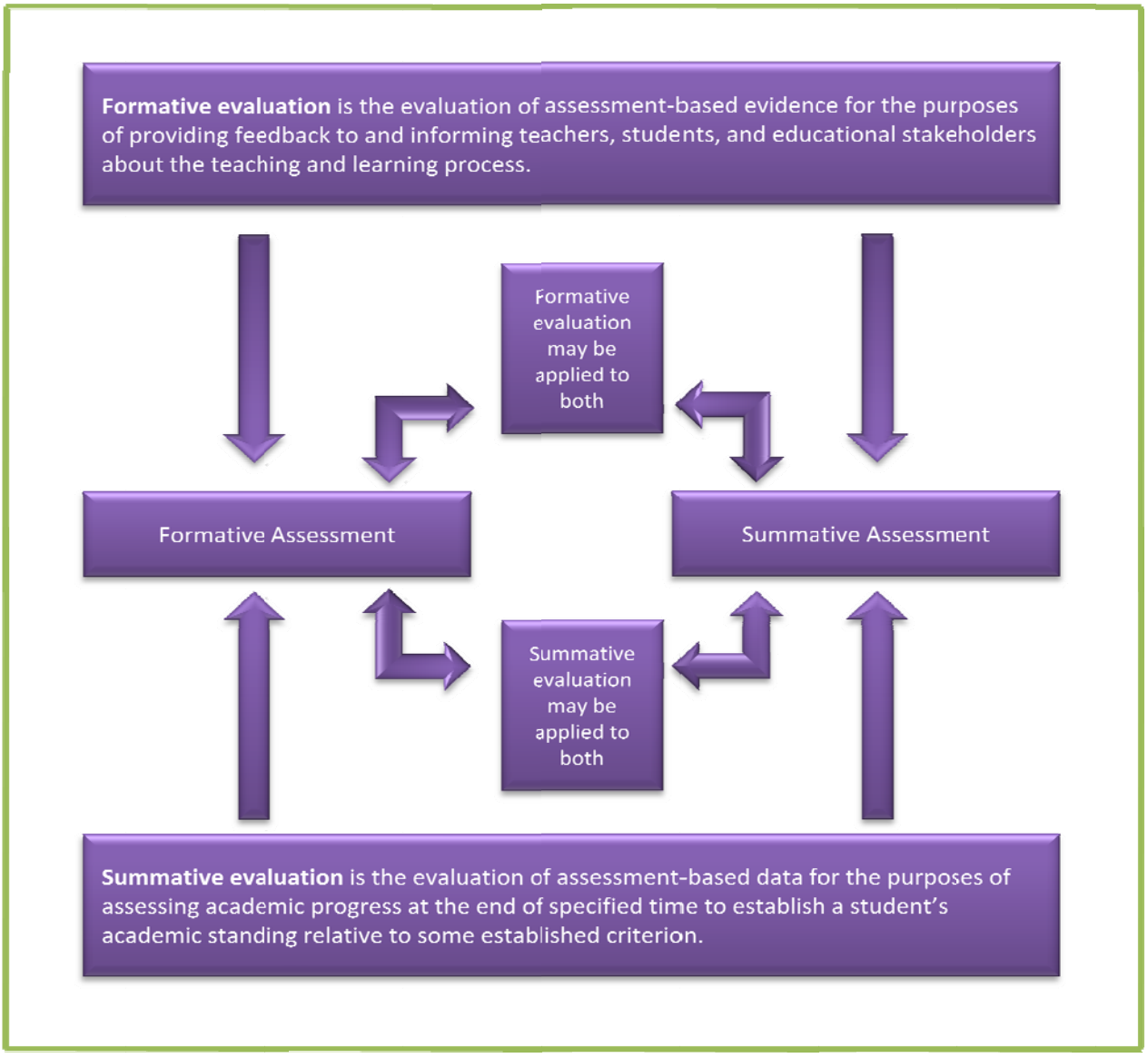

Figure 2. Practical Model of Assessment and Evaluation System

Formative Evaluation and Academic Achievement Research

In this section, we review the limited literature related to the impact of what the existing literature calls formative assessment on academic achievement. First, we examined Black and Wiliam's (1998) review of formative assessment literature. In addition, we examined more recent literature related to formative assessment and student achievement.

Black and Wiliam (1998)

In the formative assessment literature, Black and Wiliam's (1998) seminal piece is frequently cited as evidence that formative assessment does improve student achievement. In fact, one citation index that denotes all scholarly references indicates that it has been cited more than one thousand times. The Social Science Index indicates that it has been referred to in scholarly journals 194 times. This is not surprising in light of the conclusion that Black and Wiliam (1998) drew from their review of more than 250 articles related to formative assessment. They stated that the research they reviewed "shows conclusively that formative assessment does improve learning," and that the gains in student achievement were "amongst the largest ever reported" (p. 61). However, it is important to note some issues were identified with the eight research articles Black and Wiliam (1998) actually presented to support this conclusion.

The article Black and Wiliam (1998) most strongly relied on to support their conclusion was Fuch and Fuch's (1986) meta-analysis investigating the effects of formative assessment practices on student achievement. There were two primary concerns related to this article. First, of the 3,835 participants who participated in the studies reviewed, 83 percent were handicapped. The reason for this is that the review was focused on formative assessment in the context of special education; however, generalizing this to the population of students at large, as Black and Wiliam (1998) did, is inappropriate. Although Black and Wiliam do refer to the average effect size of 0.63 for non-handicapped students, which comprise the remaining 17 percent of participants in the studies reviewed, other methodological problems discussed below were of issue.

The second issue that arose from the Fuch and Fuch (1986) is they included articles that ranged from good to poor quality. In fact, of the 96 effect sizes they included in their analyses they labeled 19 as good quality, 69 fair quality, and 8 as poor quality. Because nearly 72 percent of the effect sizes included in the analyses were of fair quality, it is important to note the issues with those studies. Studies were identified as being of fair quality if they contained no more than two "less serious" methodological problems. Less serious methodological problems included: "the use of technically inadequate dependent measures, uncontrolled examiner expectancy, unchecked fidelity of treatment, the employment of inappropriate statistical unit of analysis, and inadequate teacher training" (Fuch & Fuch, 1986, p. 202). While only eight of the effect sizes that contributed to the findings were of poor quality, the combination of poor and fair quality studies in this meta-analysis accounted for 80 percent of the effect sizes included in the analyses.

Poor quality studies had at least three "less serious" methodological problems and one "serious" problem. Serious problems were defined as "unequivalent subject groups, confounded experimental treatments, and nonrandom assignment of subjects to treatments" (Fuch & Fuch, 1986, p. 202). Thus 80 percent of the effect sizes that contributed to the mean effect size of 0.70 from the 23 studies examined came from research that was methodologically unsound. In addition, it is important to note that studies that were labeled as good quality had no more than one "less serious" problem. However, if that less serious problem was of a statistical or measurement nature it casts serious doubts on the soundness of the 0.70 average effect size found by Fuch and Fuch (1986). Of greatest concern is that some of these articles had measurement and statistical issues that could directly affect the effect sizes found. While an average effect size of 0.70 is astounding, the issue of generalization to the population at large and the quality of the research reviewed creates serious problems for using this article to conclusively show that formative assessment improves academic achievement in general.

Black and Wiliam also referred to Fontana and Fernandez's (1994) research to support their conclusion. This study included 25 Portuguese teachers and 246 students from two age groups 8 and 9-year-olds as well as 10 to 14-year-olds. Teachers were trained to support students' engagement in daily self-assessment to improve math performance. The 20 control group teachers were engaged in another professional development course. Only the 8 and 9-year olds' math scores showed significant improvement when compared to the control group. The authors argued that the pre-test was too simplistic to show true difference in gains between the control and treatment groups of the older children. However, the authors did not acknowledge that the lack

of statistical difference between the control and treatment groups in the older students may also have resulted from the impact of the control group's professional development on math performance. Furthermore, it is difficult to come to conclusive decisions about the effectiveness of all formative evaluation based on a study of 25 Portuguese teachers using only self-assessments in Math with 8 to 14-year olds.

Another study that Black and Wiliam (1998) used to support their conclusion was penned by Whiting, Van Burgh, and Render (1995). In this article, an impressive seven thousand students and eighteen years of information were reviewed. The primary issue with this article is that only one teacher was studied. Although this teacher did utilize formative assessment and was compared to another teacher who did not use formative assessment, it is difficult to parse out formative assessment effects from teacher effects.

Martinez and Martinez (1992) was another study used to support the conclusion that formative assessment improves student achievement. Martinez and Martinez (1992) utilized a two by two experimental design in which two groups were taught by a novice teacher and the remaining two were taught by an expert teacher. Each teacher taught one class in which the students took only one test per chapter, and the other class took three tests per chapter. The total sample size consisted of 120 college algebra students, which resulted in small numbers of participants in each of the four sub-groups (less than 30 students each). Results indicated that the only statistically significant differences in achievement were seen between the control group (one test per chapter) and the treatment group (three tests per chapter) in the novice teacher group. The authors concluded that frequent assessment is more important in novice teachers' classrooms. This, again, does not account for teacher affects as only two teachers were investigated. Moreover, this study looked at the importance of frequency of assessment. No information was given as to the nature of the assessment or of feedback provided from the assessment, both of which Black and Wiliam (1998) use to define formative assessment. The use of this study to support the conclusion that formative assessment improves student achievement may also be inappropriate.

Black and Wiliam (1998) also reviewed an article by Bergan, Sladeczek, Schwarz, and Smith (1991). Although this study was well done, formative assessment was embedded in a larger program of measurement and planning systems (MAPS). MAPS provided teachers with assessment information and learning activities that reflected validated learning sequences. Formative assessment was an integral part of MAPS, but so were the learning activities. Therefore, it is unclear whether only formative assessment would have lead to similar significant gains in achievement if training and practice failed to include those specified learning activities. Furthermore, the population of interest in this study was impoverished kindergarteners. Due to these contextual and demographic issues, the use of this study to come to a conclusive decision about the impact of formative assessment on general student achievement is questionable.

The other three articles used by Black and Wiliam (1998) to conclude that formative assessment does improve student learning had similar problems. For example, Frederiksen and White (1997) used a treatment group that engaged in peer groups structured to promote reflection on assessments with both peer assessment of presentations to the class as well as self-assessments was compared on academic outcomes to a control group. The control group only spent time on general discussion of the learning module. However, no information about those "general discussions" was provided, nor was information provided about whether students in the control group stayed on task. Furthermore, the design of this study indicates the treatment group would have received more academic attention from the instructor. Schunk (1996) examined the effects of fourth grade math students' learning goal orientations and self-evaluations. In addition to being limited to fourth graders and self-evaluation, the sample sizes for the two studies conducted was also minimal (less than 45 students, who were then divided into smaller groupings). Although this study was well done, it does not provide support for formative assessment in general. Ross (2006) also cited Schunk (1996) as providing evidence to support the contribution of self-assessment to student achievement. The authors do not argue that this study does not support the value of formative assessment; however, neither Schunk (1996) nor the other articles cited by Black and Wiliam or Ross are enough to support Black and Wiliam's (1998) conclusion that formative assessment, or in this case self-assessment, significantly improves academic success. Finally, the Butler (1988) article presented a small sample size (48) of 11-year-old Israeli students engaging in tasks that were not presented by the teacher and that were not curriculum-based.

Collectively, the eight articles that Black and Wiliam (1998) presented to *conclusively* show that formative assessment significantly improves student achievement do not support such a conclusion. A more appropriate conclusion may have been that more research needed to be done. However, since the publication of Black and Wiliam's (1998) review, limited research has been completed to investigate the impact formative assessment has on educational outcomes. As a result, the need for empirical evidence supporting the impact of formative assessment on academic achievement still exists.

Recent Research

Nine more recent articles were also reviewed. Although these articles do provide further support for formative assessment in a fragmented fashion, methodological issues similar to those in the articles included in Black and Wiliam's (1998) review remained problematic in more current studies. For example, Thompson, Paek, Goe, and Ponte (2004) examined differences in student achievement levels in students of teachers with high and low engagement in the California Formative Assessment Support System for Teachers (CFASST) within the Beginning Teacher Support and Assessment program. Student achievement was measured on six subtests of academic achievement in California (CAT-6 Math, Reading, Language Arts, and Spelling as well as well as standardized test scores for Math and English Language Arts). No differences were found between high and low engagement of teachers, which may have resulted from the self-report nature of engagement data. Students of the teachers who participated in CFASST training did see significant gains on their CAT-6 Math, Reading, Language Arts, and Spelling scores.

The effect sizes ranged from .03 to .40, and while effect sizes are measures of practical significance it is important to note that the increases in mean performances from pre- to post tests were 3.8 on Math, 2.04 on Reading, 1.64 on Language, and 3.14 on Spelling. In light of the fact that the possible test scores ranged from 1 to 100, the practical importance of raising a score by less than four points is questionable especially when one considers the time and money dedicated to this professional development.

Wininger (2005) examined the impact of summative formative assessments on second administrations of an Educational Psychology exam. In this study, the treatment group consisted of 34 students in the researcher's Educational Psychology course. This group received feedback from the instructor and classmates, and students were guided to self-evaluate their performance. The control group consisted of 37 participants, also enrolled in the researcher's course. These students received a copy of their exam and information as to what questions they had missed, but they received no other feedback or guidance for self-evaluation. Upon the second administration of the initial test from which the students in the treatment group received feedback, the treatment group significantly outperformed the control group. The treatment group gained nine points from their initial attempt at the exam, and the control group gained only two points. In addition, Eta-squared indicated that scores from the initial examination accounted for 39 percent of the variance in the second administration of the same examination, and that the formative summative assessment treatment accounted for an additional 25 percent of the variance.

Although this study does provide support for formative assessment, a few methodological issues must be discussed. First, the small sample size results in less precise hypothesis testing and various properties of the population such as the mean used in this study to detect difference between the control and treatment group. Also of concern is a researcher bias that may have manifested in the study from a researcher using his students.

Wiliam, Lee, Harrison, and Black (2004) explored the impact of 24 teachers' use of formative assessment after a six-month training period. While the results are promising, the authors themselves noted serious issues related to generalizability of the findings. First, they stated that due to the fact that each of their results reflects a separate "mini-experiment, care needs to be taken in drawing any general conclusions about the net effect of the adoption of formative assessment" (p. 60). They further note that the method of comparison was not the same in each "mini-experiment". For example, in one study they compared students' performance to the same teachers previous group of students from the preceding academic year. In another comparison, Wiliam and his colleagues compared the performance of one teacher's students in two separate classes. These inconsistencies within the research design of this study led the authors to conclude that the quantitative evidence they provided was "difficult to interpret" (Wiliam et al., 2004; p. 62). Black and Wiliam (2003) later used this evidence to state that they could not " be sure that it was the increased emphasis on formative assessment that was

responsible for this improvement in students' scores, but this does seem a most reasonable interpretation" (p. 631). However, this contradiction in the various interpretations provided by these authors raises questions as to what evidence was identified to warrant altering their conclusion from not generalizable to generalizable.

Ruiz-Primo and Furtak (2006) found that students in classrooms where teachers engaged in assessment discussions performed significantly higher on embedded assessments and post-tests. Assessment discussions were defined as a four-stage process in which the teacher asks a question, the student responds, the teacher recognizes the response, and then uses the information collected for student learning. While these explorative results are promising, there are some issues that prevent generalizing the findings beyond the participants of the study due to the limited sample size of four.

Since the publication of Black and Wiliam's (1998) review of formative assessment, minimal scientific research on the impact of formative assessment on student achievement has been completed in the traditional classroom. However, formative assessment has been researched somewhat more thoroughly in the educational technology literature. For example, Sly (1999) investigated the influence of practice tests as formative assessment to improve student performance on computer-managed learning assessments. More specifically, Sly (1999) hypothesized that students who selected to take practice tests would outperform students who did not select to take practice tests on the first and second unit exams in a first year college Economics course. The students who selected to take practice tests did significantly outperform those who did not take practices tests on both unit exams one and two.

While Sly's (1999) results provide support for the impact formative assessment may have on achievement, this study also suffered from methodological issues. The primary issue with this study is the self-selection of participants to treatment or control groups. This is a problem because students who self-selected to take practice tests may be systematically different from those students who do not select to take practice tests. Although Sly did discuss this issue, there were no design efforts implemented to control for self-selection, through the use of instruments that measure constructs that lead to self-selection such as motivation, self-regulation, or grades prior to the use of formative assessment. In addition, while the students who selected to take practice tests did significantly outperform students on unit exams one and two, they did so by only five and four points respectively.

In another Web-based study, Henly (2003) studied the impact of Web-based formative assessment on student learning in a learning unit about metabolism and nutrition. She found that overall students in the top ten percent of the class accessed formative assessment twice as often as students in the bottom ten percent of the class. While this does reflect a significant difference in usage of formative assessments, it suffers from the same self-selection issue as Sly's (1999) study. The group that used formative assessment twice as often and ranked in the top ten percent of their class was a systematically different group from the bottom ten percent of the class who rarely accessed the formative assessment. Similar to Sly (1999), this study would have been improved by controlling for factors such as motivation, self-regulation, and prior performance. Further, in most school systems the current trend is to use formative assessments for the lowest performing students. The Sly (1999) and Henly (2003) studies have based their conclusion of the impact of formative assessments on the higher performing students, with limited evidence of their utility for these lower performing students.

Buchanan (2000) also examined the influence of Web-based formative assessment on an undergraduate introductory psychology module exam. When controlling for classroom attendance, he found that students who engaged in voluntary Web-based formative assessments significantly outperformed students who did not participate in Web-based formative assessments. However, the effect size for this difference was very small at .03. In light of the issue of self-selection in this study and the small effect size, further research with greater controls is warranted.

Wang (2007) conducted an assessment of the impact of the Formative Assessment Module of the Web-based Assessment and Test Analysis System (FAM-WATA). FAM-WATA is a multiple-choice Web-based formative assessment module containing six formative strategies: repeat the test, correct answers not given, query scores, ask questions, monitor answer history, and pass and reward. While this study showed that students who experienced FAM-WATA showed significant gains in understanding, it was compared to two other types of formative assessments without any control group to assess expected gains in performance due to instructional effects for all students. All forms of formative assessment resulted in significant student gains, with the FAM-WATA group outperforming the normal

Web-based formative assessment and the pen and paper formative assessment groups. Because the gains these groups experienced with formative assessment were not compared to the gains of students in a control group without formative assessments, it is not feasible to extrapolate that use of formative assessments was more beneficial than instruction only.

Velan, Rakesh, Mark, and Wakefield (2002) examined the use of Web-based self-assessments in a Pathology course. More specifically, the researchers hypothesized that students would do better on their third attempt at the Web-based self-assessment when compared to the first attempt. While significant improvement was seen from the first to the third attempts on the assessment, this study also had a few methodological issues. First, the sample size was very small consisting of only 44 students. Second, there was no control group. Third, the students took the same test each time, and each time they received feedback on their responses. Because the students took the same exam, it is impossible to tell whether the students gained greater understanding of the material or if they only gained expertise in taking that particular test.

Conclusion

Stiggins (2002) stated that "if we are finally to connect assessment to school improvement in meaningful ways, we must come to see assessment through new eyes" (p. 758). The purpose of this article is to provide terminology that clarifies the nomenclature related to formative and summative assessment as well as evaluation, and to highlight the need for further research with regard to formative assessment and evaluation needed to establish best practices. Thus, our goal is to provide, as Stiggins poetically stated, new eyes through which to view formative assessment and evaluation.

The research discussed in the Black and Wiliam's (1998) review and the other research discussed here does provide some support for the impact of formative assessment on student achievement. However, it provides greater support for the need to conduct research in which more efficient methodologies and designs will lead to more conclusive results and understanding of the impact of formative assessment and evaluation on student achievement. In the current NCLB era, formative assessment has been touted as an excellent means of improving student performance, in particular the achievement of lower performing students.

As a result, this is a time of both great potential and vulnerability due to the limited existence of scholarship demonstrating that students' achievement increases as a result of use of formative assessments. There is potential for the development of sound evaluation practices and statistical methodology that result in formative assessment and evaluation practices that produce powerful and positive changes in student achievement. This is also a time of great vulnerability due to the number of unproven practices that may result in the loss of time and money as well as the maximum utilization of what may be a valuable tool for improvement of educational outcomes. Thus, we do not argue that formative evaluation is unimportant, only that limited empirical evidence exists to support the "best practices" for formative evaluation. In particular, limited evidence investigates the group that may benefit the most from formative evaluation, low performing students.

An extension of this manuscript will be to develop clear more efficient research designs for evaluating the impact of formative assessment and evaluation. Too often the mere act of disaggregating student achievement data, completing a "prediction model" with a set of data, or the generation of frequency tables is represented as analyzing student achievement. Each of these statistical methods may be appropriate in various contexts, however, the assessment of student achievement and the impact of formative assessments is predicated on the type of assessment, its design, and psychometric properties. A manuscript that examines use of formative assessments, outlines appropriate methodologies, and statistical techniques is being completed to facilitate both improvement of instructional practices as well as to provide a guide for researchers who may be interested in conducting scientific studies on the impact of formative assessments.

Stiggins (2005) stated that "to use assessment productively to help achieve maximum student success, certain conditions need to be satisfied" (p. 4). The authors agree with this statement, but also posit two conditions that were not stated by Stiggins. First, a clear and shared lexicon must be established and shared among all educational stakeholders to lead to more productive communication among teachers, researchers, policy makers, parents, and students. Finally and most importantly, a sound research-validated framework for best practices in formative assessment and formative evaluation must be established to ensure maximum benefits for all those involved.

References

Bell, B., & Cowie, B. (2000). The characteristics of formative assessment in science education. *Science Education*, *85*, 536–553.

Bergan, J. R., Sladeczek, I. E., Schwarz, R. D., & Smith, A. N. (1991). Effects of a measurement and planning system on kindergartners' cognitive development and educational programming. *American Educational Research Journal, 28*, 683-714.

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

Black, P., & William, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal, 29*(5), 623-637.

Bloom, B.S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Taylor (Ed.), *Educational evaluation: New roles, new means: The 68th yearbook of the National Society for the Study of Evaluation, Part II* (pp. 26-50). Chicago: University of Chicago Press.

Buchanan, T. (2000). The efficacy of a World-Wide Web mediated formative assessment. *Journal of Computer Assisted Learning, 16*, 193-200.

Butler, R. (1988). Enhancing and undermining intrinsic motivation; the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*, 1-14.

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

Fontana, D., & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology, 64*, 407-417

Frederiksen, J. R., & White, B. J. (1997). Reflective assessment of students' research within an inquiry-based middle school science curriculum. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

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

Henly, D. C. (2003). Use of Web-based formative assessment to support student learning in a metabolism/nutrition unit. *European Journal of Dental Education, 7*, 116-122.

Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing, 21*(3), 335-359.

Martinez, J. G. R., & Martinez, N. C. (1992). Re-examining repeated testing and teacher effects in a remedial mathematics course. *British Journal of Educational Psychology, 62*, 356-363.

Melmer, R., Burmaster, E., & James, T. K. (2008). *Attributes of effective formative assessment.* Washington, DC: Council of Chief State School Officers. Retrieved October 7, 2008, from http://www.ccsso.org/publications/details.cfm?PublicationID=362

Perie, M., Marion, S., & Gong, B. (2007). The role of interim assessments in a comprehensive assessment system: A policy brief. Retrieved October 1, 2008 from http://www.nciea.org/publications/PolicyBriefFINAL.pdf

Popham, W. J. (October 2006). *Defining and enhancing formative assessment.* Paper presented at the Annual Large-Scale Assessment Conference, Council of Chief State School Officers, San Francisco, CA.

Popham, W. J. (2008). *Transformative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research and Evaluation, 11*(10). Retrieved January 31, 2009 from http://pareonline.net/getvn.asp?v=11&n=10

Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment, 11*, 205-235.

Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal, 33*, 359-382.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, and M. Scriven (Eds.), *Perspectives of curriculum evaluation, Volume I* (pp. 39-83). Chicago, IL: Rand McNally.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Sly, L. (1999). Practice tests as formative assessment improve student performance on computer-managed learning assessments. *Assessment and Evaluation in Higher Education, 24*(3), 339-343.

Stiggins, R. J. (2002). Assessment Crisis: The absence of assessment FOR learning. *Phi Delta Kappan, 83*(10), 758-765.

Stiggins, R. J. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan, 87*(4). Retrieved February 1, 2009 from http://www.pdkintl.org/kappan/k_v87/k0512sti.htm

Thompson, M., Paek, P., Goe, L., & Ponte, E. (2004). *Study of the California formative assessment and support system for teachers: Relationship of BTSA/CFASST and student achievement.* Princeton, NJ: Educational Testing Service.

Velan, G. M., Rakesh, K. K., Mark, D., & Wakefield, D. (2002). Web-based self-assessments in Pathology with Questionmark Perception. *Pathology, 34*, 282-284.

Wang, T. H. (2007). What strategies are effective for formative assessment in an e-learning environment? *Journal of Computer Assisted Learning, 23*, 171-186.

Whiting, B., Van Burgh, J. W., & Render, G F. (1995). *Mastery learning in the classroom.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*, 49-65.

Wininger, R. S. (2005). Using your tests to teach: Formative summative assessment. *Teaching Psychology, 32*(2), 164-166.

Citation

Dunn, Karee E and Mulvenon, Sean W. (2009). A Critical Review of Research on Formative Assessments: The Limited Scientific Evidence of the Impact of Formative Assessments in Education. *Practical Assessment Research & Evaluation*, 14(7). Available online: http://pareonline.net/getvn.asp?v=14&n=7

Corresponding Author

Karee E. Dunn
248 Grad Ed
College of Education
University of Arkansas
Fayetteville, AR 72701
Phone: (479) 575-5593
Email: kedunn [at] uark.edu