A Guide to the Formative Assessment Delivery System (FADS)

(Ver. 1.1, June 2009)

Mark Wilson, Kathleen Scalise, Andrew Galpern, & Yi-Hung Lin

Berkeley Evaluation and Assessment Research (BEAR) Report, UC Berkeley, Berkeley, California
## FADS Guide: Table of Contents:

- Introduction to the FADS guide: Page 3
- Introduction to the FADS project: Page 4
- Formative Assessment: Page 5
- The BEAR Assessment System: Page 8
- FADS Item Design and Signature Item Types: Page 20
- Contacts and resources: Page 23
Introduction to the FADS Guide

Welcome! If you’re reading this, then you’re probably a teacher or researcher or someone interested in learning more about FADS. This guide is designed for you, and we hope you find it useful.

This guide will introduce you to the key components of FADS and show you how to integrate FADS with everyday classroom teaching.

In a nutshell, FADS is a computerized assessment delivery system that allows teachers to more easily design and administer assessments, collect meaningful data about what your students know, and respond to that data with appropriate interventions in the classroom.

FADS is a teacher-driven and teacher-tested system. We believe that the most useful tools in the classroom must be useful and practical for real-life classroom teachers in everyday classrooms. It must be easy to use and the information it provides needs to be meaningful and useful to teachers.

Note: Updates to this guide will be made regularly throughout the FADS project as designs are implemented. Check the FADS website for the most current version of this guide:

http://bearcenter.berkeley.edu/projects/FADSweb/ProfessionalDevelopment.php
Introduction to FADS

With the interest of helping teachers improve their instructional practices and enhancing student learning, FADS (Formative Assessment Delivery System) is a computerized system that will allow classroom teachers to design, develop, and deliver formative assessments and to monitor and report student progress within an interpretive context. This online accessible system will allow teachers to accurately diagnose students’ comprehension and learning needs by providing real-time assessment, logging, analysis, feedback, and reporting. The current five-year FADS project, funded by the National Science Foundation, is focused on designing activities deriving from middle school mathematics and science curricula aligned with state and national standards.

A key feature of this project is the centrality of teacher expertise. We collaborate extensively with teachers to learn how technology can make a difference in the classroom. Teachers will guide the design of (1) the way feedback is provided to students during instructional uses of assessment activities (e.g., embedded assessment), (2) the reports that will be provided to teachers after students complete assessments, and (3) the professional development materials about how to effectively gather and use formative assessment data.1

1 From the FADS website
Formative Assessment

The benefits of using high-quality formative assessment data to inform teachers’ day-to-day and longer-term classroom decisions have been rigorously established in the landmark meta-analysis by Black & Wiliam (1998). Researchers have come a long way towards agreeing on what “good” assessment looks like, with clear connections between what is taught and what is assessed (Pellegrino, Chudowsky, & Glaser, 2001). The landscape is less well defined in describing how teachers might use assessment data to improve student learning.\(^2\)

All teachers are familiar with assessment because assessment, in one form or another, has been a part of education from the beginning. More recently, the No Child Left Behind (NCLB 2002) legislation, has forced teachers, for better and for worse, to become very aware of assessment and testing and its complicated (and sometimes contradictory) relationship to teaching and learning, student performance and accountability.

Formative and summative assessment differ in a number of important ways including when it occurs, what form it takes, and its purpose.

Formative assessment occurs in the midst of instruction, and is used to guide teaching and learning.

Summative assessment occurs at the end of some unit of instruction, and is generally used to report and compare performances.

Formative assessment is (ideally) embedded in instruction and indistinguishable from it. Teaching, learning, and formative assessment can occur simultaneously.

Summative assessment is often distinct from instruction, and often has a different look and feel from the instruction that preceded it. (e.g., standardized “bubble” tests)

The purpose of formative assessment is to measure student understanding so that teachers can respond and adjust their teaching plans and methods to best meet student needs. Formative assessment also gives students the information they need to improve their performance.

The purpose of summative assessment is to measure student understanding after some discreet unit of instruction (e.g., a chapter, a unit, a semester, etc.)

\(^2\) from the AERA 2009
FADS takes a very broad view of assessment. More than just converted paper and pencil items to an online environment, FADS provides greater flexibility for teachers and a greater range of student response types and question formats including multiple choice, written (typed) responses, and some student performances/behaviors with online tools. More information about item formats can be found in the section: FADS Item Design and Signature Item Types.

**Black and Wiliam**

In 1998, Paul Black and Dylan Wiliam concluded an analysis of over 500 previous studies, to answer key questions about the effectiveness of formative assessment. They found significant gains in student achievement in classrooms that engaged in the best practices of formative assessment.


Black and Wiliam describe assessment in this way:

We use the general term assessment to refer to all those activities undertaken by teachers -- and by their students in assessing themselves -- that provide information to be used as feedback to modify teaching and learning activities. Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching to meet student needs.

Some of the key ideas in their research include:

• The results of formative assessments should be used to inform both students and teachers, so that they can each make changes to the teaching and learning process

• Feedback should be given in a timely matter, within instruction, and offer the student specific advice and suggestions about how to improve their work.

• Students could and should be involved in the process of assessment by developing and understanding the rubrics for achievement

• The use of formative assessment has strong positive influences on self-esteem and motivation, which are both linked to student achievement.

3 Black and Wiliam’s commentary is available online here [http://www.pdkintl.org/kappan/kbla9810.htm](http://www.pdkintl.org/kappan/kbla9810.htm)
Formative Assessment in FADS

The FADS project is building a system that is consistent with the best formative assessment practices (Black & Wiliam, 1998) and can provide teachers with the assessment design and reporting tools they need to make formative assessment decisions. The system is flexible enough to allow both teacher and student input, and the reporting options available will help teachers understand both individual and group performance, as well as provide specific advice for individual students, so that teachers may make adjustments to the pace and content in their classes.

Other reporting options we envision for FADS include automated and timely feedback for students. By offering real-time, or near real-time feedback to student responses, we hope to capitalize on the smallest “teachable moments” by providing students with additional assessment and/or instructional tasks, as well as the specific feedback they will need to improve, all delivered within the same system, integrating assessment, instruction, and feedback in a continuous and iterative process that is controlled and managed by classroom teachers, and engineered to drive student performance.
The BEAR Assessment System and the Four Building Blocks

The BEAR Assessment System: A Brief Summary for the Classroom Context

The Bear Assessment System (BAS; Wilson & Sloan, 2000) is an integrated approach to developing assessments that provide meaningful interpretations of student work relative to the cognitive and developmental goals of a curriculum. It is grounded by four key principles guiding assessment development and includes four building blocks (each associated with one of the principles; Wilson, 2005) that are tools for constructing meaningful assessments aligned with curricular goals and instructional activities. These principles are:

- Assessment should be based on a developmental perspective of student learning.
- What is taught and what is assessed must be clearly aligned.
- Teachers are the managers and users of assessment data.
- Classroom assessment must uphold sound standards of validity and reliability.

These four principles also relate to the Assessment Triangle developed by the National Research Council Committee on the Foundations of Assessment and published in their report, Knowing What Students Know (NRC, 2001). The Assessment Triangle, shown in Figure 1, is a model of the essential connections and dependencies present in a coherent and useful assessment system. In this triangle, assessment activities (the observation vertex) must be aligned with the knowledge and cognitive processes (the cognition vertex) one wishes to affect through the instructional process, and the scoring and interpretation of student work (the interpretation vertex) must reflect measures of the same knowledge and cognitive processes. Meaningful connections among the three vertices, cognition, observation, and interpretation, are deemed essential for assessment to have a positive impact on learning. We refer to this whole process as construct modeling.

Four Building Blocks

The BEAR Assessment System includes four building blocks for constructing quality assessments: Construct Maps, Items Design, the Outcome Space, and the Measurement Model. These building blocks map to the NRC Assessment Triangle, as shown in Figure 3 below, and operationalize the BEAR Assessment System principles. (does the BEAR four building blocks really related to the NRC triangle? How to explain?)

---

4 This section is adapted from an excellent earlier paper written by Cathleen Kennedy (Kennedy 2005) available at [http://bearcenter.berkeley.edu/publications/BAS_Summary.pdf](http://bearcenter.berkeley.edu/publications/BAS_Summary.pdf).
Guide to the Formative Assessment Delivery System (FADS)

Assessment Triangle

Observation  Interpretation

Cognition

Figure 1. The National Research Council's assessment triangle.

BEAR Assessment Building Blocks

Items Design  Outcome Space & Measurement Model

Construct Maps

Figure 2. BEAR Assessment System building blocks.
FADS is built with the same principles found in the BEAR Assessment System, and uses the same four building blocks within an online delivery system.

The four building blocks are:

1. The Construct (also commonly known as a learning progression)
2. The Items Design
3. The Outcome Space (also commonly known as a scoring guide or rubric)
4. The Measurement Model

These four building blocks are described in greater detail below.

**A construct is what you want to measure**

A *construct map*, which defines a latent variable or construct, is used to represent a cognitive theory of learning consistent with a developmental perspective. This building block is grounded in the principle that assessments are to be designed with a developmental view of student learning. This means that the underlying purpose of assessment is to determine how students are progressing from less expert to more expert in the domain of interest, rather than limiting the use of assessment to measure competence after learning activities have been completed. Generally, we want to describe a continuum of qualitatively different levels of knowledge from a relatively naïve level to a more expert level. The knowledge, skill, or ability one wishes to measure is referred to as the construct, while the definitions of what it means to have more or less of the construct, in a cognitive sense, is referred to as the construct map. A construct map serves as a mechanism for defining and representing what students know and can do at several levels; it facilitates the alignment of responses on student work with the cognitive goals of the curriculum. An example of a *generic* construct map is shown in Figure 3. In this example, three qualitatively different levels of knowledge are defined.

---

5 The BAS can be applied to assessment situations in which single or multiple continua are the underlying form of the construct. It is not designed for other underlying forms, such as latent classes.
Figure 3. A generic Construct Map describing how students are expected to perform at three levels of understanding on the construct.
The items design are the questions or tasks that are used to inspire responses from students

The items design building block is a framework for designing tasks to elicit specific kinds of evidence about student knowledge, as described in one or more construct maps. Traditionally, the items design has most often taken the form of paper and pencil tests. The FADS project is being built to expand the possibilities for the design of items, including fixed responses (multiple choice) and more open-ended constructed response types where students can write, speak, or interact with online tools when responding to a question. In this way, teachers will have more flexibility to design and test items for their classroom. Some examples of more innovative item types can be found in [this section or this page] (?

The guiding principle is that assessment should be seamlessly integrated into the instructional activities of a course. That is, assessment is not merely tacked on at the end of instructional units, but is embedded in normal classroom activity and may even be, from the student’s point of view, indistinguishable from instruction. This necessitates unambiguous linkages between assessment activities and curricular content so that assessment results can be used to improve teaching and learning activities in an ongoing manner. Items are written with the intention of producing evidence of specified levels of understanding along a construct. In some cases, for example, in embedded assessments, it may be desirable for a collection of items in a particular instrument to focus on only one or two levels at a time; in other cases, such as in the end of unit summative assessments, one may want a collection of items that cover the complete range of the construct. In addition, items may be designed to measure multiple aspects of knowledge in a single comprehensive response, thereby producing more information from an amount of student effort that might otherwise provide evidence of only one aspect of knowledge (and then multiple pieces of information may be mapped onto multiple construct maps).

The outcome space is a plan to make sense of student responses to the items

The outcome space describes in detail the qualitatively different levels of responses associated with the construct map for a particular prompt or stimulus. This building block operationalizes the principle that teachers are to be the primary managers of assessment in the classroom. To accomplish this, they must have not only the data that they need to assess student learning, but also the skill set required to use that data effectively. This implies a data-driven approach to teaching, in which teachers use assessment evidence to draw inferences about student knowledge and understanding. This evidence can then suggest actions that teachers and students can take to improve learning outcomes. In addition, teacher judgment becomes a valued and integral part of the assessment feedback loop. The purpose of the outcome space, then, is to facilitate identification of student responses corresponding to a particular level on a construct; teachers use the outcome space to assign scores to student work. The outcome space and the associated construct map become the evidentiary foundation for teachers to use on a daily basis in their classrooms, on both formal assessments and informal instructional contexts.

We often organize moderation sessions in which teachers discuss with one
another how they might score individual student work products. One objective is to improve consistency between scorers and to help teachers become more consistent in their own scoring. Perhaps more importantly, this activity also helps teachers internalize the developmental goals of the curriculum as they focus attention on student locations along the construct continua and reflect on a) what it means to be located at one level or another, and b) what sort of evidence is required to conclude that one knows where a student is located on the construct map.

The measurement model is the plan to connect data from student responses with inferences about student performance.

The final building block of the BEAR Assessment System is the measurement model, which defines how inferences about student understandings are to be drawn from the scores. The principle underlying this building block is that classroom assessment should adhere to sound quality standards, providing evidence for both validity and reliability.

Of course, the classroom context is different from that of traditional standardized assessment, so the manifestations of quality may well differ. This principle helps ensure that the inferences drawn from classroom assessment activities are meaningful in their own right, and also consistent across multiple instruments. The latter is particularly useful when making longitudinal comparisons to determine change or progress. The psychometric approach we often use to model the data is a multidimensional Rasch-based item response model known as the multidimensional random coefficients multinomial logit model (MRCMLM) (Adams, Wilson & Wang, 1997). This model provides a convenient way to develop person proficiency measures and item difficulty measures using the same scale; a technique that facilitates the interpretation of student measures on the construct. An example of this alignment is shown in Figure 4. For example, we can describe what a student located at a certain level on a construct can be expected to do based upon items aligned at that level and below it on the same scale. When a person and a dichotomous item are parallel to one another on the map, the person has a 50-50 chance of responding correctly to that item. The person has a higher (than half) chance of responding correctly to items below his or her location, and a lower chance of responding correctly to items above his or her location. Similar statements can be made for specific scores on polytomous items. In the example shown in Figure 4, Student A has a 50-50 chance of responding correctly to item 3, and a higher chance of performing correctly on items 1 and 2.
The BEAR Assessment System includes software to facilitate the design of assessments and model respondent proficiencies. The software, called GradeMap (Kennedy, Wilson & Draney, 2005), produces maps to help teachers interpret student positions on the constructs of interest. One example, shown in Figure 5, is a Performance Map, which shows how a student has performed on a series of assessment activities over time. Note that the titles on the right edge of the map are brief descriptions of the levels from the construct map for the associated curriculum. In this example, the construct map for the Designing and Constructing Investigations construct differentiates (two verbs?) five levels of performance: off target, incorrect, incomplete, correct, and beyond correct. Each point on the map shows the location of the student’s most active learning at that point in time. This differs from the concept of competence in that each point represents the location on the construct where the student had a 50-50 chance of performing at that level or above.2 (where is footnote 2?) In this example, the student’s responses tended to be at the Incomplete level for activities 1 through 28, and then her responses become mostly Correct after that.
Figure 5. Performance Map for Anne Jackson produced by GradeMap showing maximum likelihood proficiency estimates on the “Designing and Conducting Investigations” construct at five time points.
Teachers can also analyze the performance of a class as a whole with an output map called the Frequency Map (which has been named a Class Map in the example). In the example shown in Figure 6, taken from a curricular unit about buoyancy, a teacher can see that after completing a formative assessment at the fourth week of instruction, about 25% of the students’ responses were below the targeted Mass level; that is, these

![Class Map for All Students]

Map of student proficiencies after 4th week.

![Class Map for All Students]

Map of student proficiencies after 7th week.

Figure 6. Frequency maps of student proficiencies at the 4th and 7th weeks of the curriculum.
students were unable to explain how changing the mass of an object affects how it might sink or float. The teacher can then compare that map with a map of student proficiencies after the seventh week to see that the class has improved overall.

Now, only a small percentage of students cannot explain how mass affects buoyancy, and well over 50% are able to explain that mass and volume work together to affect floating and sinking (i.e., at least 50% of the students had a proficiency level at or above the Mass + Volume level), which is the targeted learning objective during the seventh week of instruction.

This type of information can help teachers focus classroom activities in ways that will be most beneficial to their particular students. For more complete examples of GradeMap reports and their interpretations, refer to the *GradeMap Users Guide* (Kennedy, 2005). As these two examples demonstrate, the GradeMap software aligns student proficiencies with the constructs targeted by a curriculum. This can improve the interpretability of student work and help teachers focus on the specific needs of their students in the context of the developmental perspective of the curriculum. Assessments designed and developed within the BEAR Assessment System framework employ these principles and building blocks.

This approach helps ensure that appropriate evidence is produced to draw reliable inferences about the student proficiencies of interest, and that those inferences can be interpreted in a straightforward way that is meaningful for teachers, students, and other educational stakeholders.

The teacher can then compare that map with a map of student proficiencies after the seventh week to see that the class has improved overall. Now, only a small percentage of students cannot explain how mass affects buoyancy, and well over 50% are able to explain that mass and volume work together to affect floating and sinking (i.e., at least 50% of the students had a proficiency level at or above the Mass + Volume level), which is the targeted learning objective during the seventh week of instruction. This type of information can help teachers focus classroom activities in ways that will be most beneficial to their particular students. For more complete examples of GradeMap reports and their interpretations, refer to the *GradeMap Users Guide* (Kennedy, 2005). As these two examples demonstrate, the GradeMap software (where is the footnote 3?) aligns student proficiencies with the constructs targeted by a curriculum. This can improve the interpretability of student work and help teachers focus on the specific needs of their students in the context of the developmental perspective of the curriculum.
Assessments designed and developed within the BEAR Assessment System framework employ these principles and building blocks. This approach helps ensure that appropriate evidence is produced to draw reliable inferences about the student proficiencies of interest, and that those inferences can be interpreted in a straightforward way that is meaningful for teachers, students, and other educational stakeholders.
References


Items Design and Signature Item Types

At this point in the FADS development process, classroom teachers have contributed to the design of “signature” item types, to demonstrate the range of design possibilities for future users of the system, and to provide an item design “bank” that can be used flexibly with different content from both math and science curricula.

What you see below are screenshots of the draft item designs that were piloted in the Spring of 2009.

FADS Signature Item Types

While FADS is being designed to deliver a variety of standard item types such as multiple-choice and short-answer essay questions, two assessment content technologies that to date have had limited research in K-12 education have been selected as the core FADS signature item types: Natural user interfaces (NUI) and adaptive virtual simulations (AVS). These two types are expected to be fully enabled within the system, and offer a rich range of assessment opportunities.

The FADS system is being developed with an item pool focus. This means that all items considered to be in a single “library,” or belonging to a single construct or dimension of measurement interest, are located in a single pool, from which various instruments are created by selecting and sequencing from the pool. This facilitates both generation of a range of new instruments, and calibration of items across the possible range.

---

6 From the FADS_AERAfinaAprill09.doc
Professional Development Materials ver. 1.18/30/09 11:44 PM
Figure 3. Examples of the delivery of a NUI object in the first FADS trials, for middle school students. Using a sample container environment that can deliver computer adaptive tests (CAT), a NUI object Card Sort was presented to students. Students examined 11 food labels and sorted each label into the appropriate category of “Food” or “Not Food,” based on the information about carbohydrates, protein, fat and other substances in the food type. The activity was scored automatically, and then several short-answer questions extensions were delivered to students. The three extension items were intended to be scored by teachers, and a single teacher’s set of items for up to 120 students was found possible to score according to rubrics in about 10 minutes per question.

Figure 3. Examples of the delivery of a NUI object in the first FADS trials, for middle school students. Next, a NUI Color Transformation object was presented to students, in which students predicted a model change based on a simulation of air in a closed flask.
Figure 4. Examples of the delivery of a NUI object in the first FADS trials, for middle school students. Finally, a NUI Math Representations object was presented to students. Here students were provided with a data table with information on the trout population in one of the Great Lakes. Students adjusted points on the graph to reflect the data provided, and scores were delivered to the system. For this object, a web interface was designed so that teachers could insert any question prompt, data table and set of original data points on a graph, allowing for the possible creation of many different assessment objects.

It is also expected that testlet models will be needed to calibrate the item banks. Online content such as described here is relatively expensive to develop. Thus multiple items are usually situated in a single context, and share some of the same prompt and stimulus material, thus indicating the need of testlet models that take into account local dependencies within the testlet. One computerized testlet approach for modeling complex assessment tasks is the Iota Model (Scalise & Scalise, 2004; Scalise & Wilson, 2005). This model accommodates tasks in which multiple paths through the problem may produce similar overall scores but different interpretations of student understanding.

For example, a task may begin with some common stimulus material, such as information and a question, and then students receive subsequent questions, or probes, adaptively, depending on their responses. Different paths of questions and answers can subsequently result in unique scores or equate to the same score and interpretation, depending on expert advice and the results of data analysis. Modeling evidence-rich tasks in this way can provide more reliable and useful information about student understanding and needs than using traditional methods.

A number of other researchers have explored additional item formats that require the construction of a response but can be scored automatically, for example, Ordered Multiple Choice (Briggs, Alonzo, Schwab, & Wilson, 2006), Meaning Equivalence Methodology (Seeratan, 2006), and the Computerized Modified Multiple-Choice Testing System (Park, 2005). We have identified more than a dozen item formats that could be used to assess deep knowledge and critical thinking skills and that can be scored through automated procedures. In this project we anticipate testing these and other formats that we identify to broaden the selection of item formats available to assessment developers for computerized delivery and scoring.
Contact Information and Resources

The FADS project is a collaboration between classroom teachers and researchers at the University of California at Berkeley, The University of Michigan, and Vanderbilt University.

The FADS website can be viewed here:

http://bearcenter.berkeley.edu/projects/FADSweb/

and contacted here:

http://bearcenter.berkeley.edu/projects/FADSweb/ContactUs.php

To contact us with general comments/inquiries about FADS, please send an email to bearit@berkeley.edu

Collaborating Institutions:

Center for Highly interactive Classrooms, University of Michigan; Learning Sciences Institute, Vanderbilt University; Berkeley Evaluation & Assessment Research (BEAR) Center, University of California at Berkeley

Funded by:

National Science Foundation, Discovery Research K-12

University of Michigan Participants:
Joe Krajcik
Namsoo Shin
Joi Merritt

Vanderbilt University Participants:
Rich Lehrer

BEAR Participants:
Mark Wilson (PI)
Cathleen Kennedy
Kathleen Scalise
Amy Dray
Sevan Tutunciyar
Ana Maria Albornoz Reitze
Ya-Hui Su
Andrew Galpern
Yi-Hung Lin
Grace Kim