Diagnostic Questions: Is There Value in Just One?[1]

Caroline Wylie

Dylan Wiliam

ETS, Princeton, NJ

Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) held between April 6 to 12, 2006, in San Francisco, CA.

[1] The work reported in this paper was supported by the Institute for Education Sciences (IES) under grant number R305K040051, although the views expressed here are solely those of the authors.

Introduction

Providing teachers with better analyses of test data at regular intervals during the school year may go some way to help teachers better align their curriculum with key standards and to ensure that teachers within a school align themselves to a common curriculum. This approach may be of limited utility, however, when it comes to yielding substantial and sustainable increases in student learning since by the time the information is available the class has moved on to another topic. Such analyses may assist in improving the alignment between instruction and assessment, but the gains are likely to be small, and ethically problematic, since, in effect, they would be produced by "teaching to the test."

Furthermore, since the use of assessment undertaken for the purpose of teacher and student accountability can appear to be more related to the needs of administrators and politicians, the work involved in making use of such data is also likely to be perceived by teachers as a bureaucratic addition to their workload, rather than a better way of discharging their existing responsibilities. Finally, since teachers are not already engaged in such use of data, teachers will have to find time to integrate these new activities into existing, well-established routines, and the literature on teacher change suggests that this will be slow, if it happens at all.

An alternative is to start from the activities that teachers already undertake, and in particular the decisions that they make regularly, and work from there. This is the approach adopted by the *Using and Applying Diagnostic Items in Math and Science* (DIMS) project funded by the Institute of Education Sciences. The central idea of the DIMS project is that teachers can use the results from single, high-quality questions to support the making of rapid judgments about student understanding and make "on-the-fly" adjustments to their instruction, so as to better meet student learning needs (Leahy, Lyon, Thompson and Wiliam, 2005). Such items might be used as "range-finding" questions at the beginning of instruction to judge where to pitch the presentation, in the middle of a sequence of instruction to judge whether the class is ready to move on, or at the end of a sequence of instruction to judge where the intended material has been learned, in order to provide information about the most appropriate starting point for the next instructional episode.

This paper will contrast these two approaches to "data driven instruction," contrasting a "data-push" with a "decision-pull" approach. It will then elaborate on the key principles of the DIMS project, and in particular the utility of using information from single items to influence and direct classroom instruction.

Data-Push vs. Decision-Pull

Within the last few years, a number of vendors have started marketing systems variously described as "formative tests" or "formative assessments". These include the MAP

produced by the Northwest Evaluation Association (NWEA), the Focus on Standards™/Instructional Data Management System™ produced by ETS, *Homeroom*™ produced by Princeton Review, *Benchmark Tracker*™/*SkillWriter*™ and Stanford Learning First™ by Harcourt Assessment, and Prosper™ produced by Pearson Assessments, as well as a host of other similar systems.

Although there are important differences between these products, there are also common features. Students are assessed on a regular schedule, with the interval ranging from four to ten weeks, and analyses of student performance are shared with teachers. Teachers typically meet in pairs or groups to review and discuss the results, and make decisions based on the data regarding next steps. Depending on the system, these next steps may take the form of adjustments to the curriculum to secure better alignment with state standards (or tests that are based on the standards which of course is not at all the same thing), or the next steps may take the form of "interventions", such as arranging for students performing poorly on the assessments to receive additional instruction (in recess periods, after school, on Saturdays, or in vacation periods). Although many vendors claim that these kinds of usages allow assessment to be integrated with assessment, the extent to which the assessment is integrated with instruction is loose in two senses. The first is that these assessments operate on a basis that is closer to quality control than quality assurance. The tests are given at the end of a sequence of instruction, and where the tests reveal that the level of student achievement is less than required, the students (and usually only those students) are routed towards additional sequences of instruction—a process that might be termed "match, batch and dispatch". The second sense in which the integration of assessment with instruction is loose is that the results are generally focused more on monitoring the level of student achievement, or on diagnosing the nature of the problem in broad terms, rather than providing the teacher with guidance about what to do about it (Wiliam & Thompson, 2006). Although teachers will know *that* some action is required, the fact that the results are reported at such a general level will mean that the teacher will get little specific guidance about *what* needs to be done.

In reality, teachers hardly ever ask the question, "Do I need to remediate?" once a particular episode of instruction has ended and the class has moved on, unless the topic reappears later in the year as pre-requisite knowledge for a new topic. Thus, teachers do not have structures in their schedules that permit them to take and act on this new information, nor do they have experience with incorporating this medium-length cycle information into their planning. And, as Helmke and Schrader (1987) found, teachers who collect information about individual student achievement, but who are unable to do anything with the information, get worse results than teachers who never bother to collect such information in the first place.

We consider this cycle of assessment every six or so weeks as an example of a "data-push" system: data is pushed at the teacher with the expectation that somehow it will be used to improve or reform instruction, much along the same lines that Kevin Costner had faith that "If you build it, they will come" in the film *Field of Dreams*. Having received the data the teacher now knows that certain topics covered in the previous month for some or all students were not adequately learned, but the teacher has no information that

will help him or her in the planning of the current lesson. It will be another four weeks before that the information becomes available. Perhaps most importantly, the use of the information that is available requires teachers to add new routines to their existing practice, such as arranging for some students to receive remedial instruction, while others are provided with "extension" material.

This is why, in thinking about how assessment might be integrated with instruction, in the DIMS project, we decided to start with the decisions that might be informed by data rather than the data itself. Rather than providing unwanted answers to unasked questions (*cf*. Popper, 1976 p. 40), we decided to explore the possibilities for starting from the types of decisions that we know that teachers do make, on a regular and frequent basis.

A prime candidate is when the teacher asks herself "is the class ready to move on?" This question can play out at a variety of levels—a teacher might ask it before deciding to move on to the next activity during a lesson, or ask it before leaving a particular topic and moving on to the next. In any case, the issue of "readiness to move on" is a familiar one to teachers, and the decision not to move on will initiate routines that are also familiar: to engage in whole class remediation, to pull aside a small group or individual students for additional assistance, to construct alternative learning opportunities that will assist students in their learning, and so forth, so that the teacher can later reassess the situation and proceed in the learning sequence.

However, observation of practice in classrooms suggests that the evidence-base for these decisions is often very shallow. Typically, a teacher will ask a question of the class, see a handful of "the usual suspects" raise their hands, and select one of this handful to respond. If the teacher gets a correct response from this one student, she tends to assume that all other students in the class share the understanding, and move on.

So classroom assessment as conceived by the DIMS project begins with the premise that teachers need information that they can use "on-the-fly" to guide instructional decision-making. Starting from this point, then the next question to consider is what is the type of information that teachers can and should process during the course of a lesson?

Characteristics of "Decision-Pull" Questions

The DIMS project is designed to assist teachers with instructional decision-making from the "decision-pull" perspective. The questions that teachers use in the project have the following characteristics:

- Designed for easy collection of information;
- Incorrect responses assist the teacher diagnose what students do not understand, and, ideally, provide ideas about what to do about this.

- Correct responses support a reasonable inference that students understand the concept being assessed;

In order to overcome one of the problems of classroom questioning mentioned in the previous section, the first characteristic of questions used in the DIMS project is that they are designed so that the teacher can collect information from every student in the class easily and quickly. Most (although not all) questions are in a multiple-choice format, although many of the items have multiple correct answers. The advantage of using items with multiple correct answers is that the outcome space is much larger than with traditional multiple-choice items. For example, if the item asks students which of a list of six things is living, there are 2^6 possible outcomes, only one of which is correct, so that a student's chance of guessing correctly is less than 2%, compared with approximately 17% with a traditional multiple-choice item. In this context, it is also worth noting that most of the commercially available electronic "clickers" that allow students to beam their responses to the teacher's computer via infra-red or radio-frequency messages only allow a single response to an item, and so cannot provide such rich information as the card-based system.

In the classrooms participating in the DIMS project, every student in the class has a set of cards with the letters A to H printed on each card. The letter on each card is in a sufficiently large font to be seen from the front of the classroom. Typically, the question is displayed using an overhead projector, students individually select their response(s) and then on cue (and not before!) hold up one or more cards for the teacher to see.

The teacher is able to scan the room to get a sense of the range of responses. He or she can then make an informed instructional decision about what to do next. If all students indicate correct solutions, then the teacher can move on, reasonably confident that the class has understood the intended point. If no students answer correctly, then she might decide to re-teach the concept using another approach. Generally, however, there will be a range of correct and incorrect responses, which provides the teacher with a variety of options. She could ask a student with an incorrect answer to talk about his or her reasoning; select a student with the correct answer to explain to the class; decide not to reveal the correct answer but rather to proceed with instruction, and return to the question again at the end of the lesson, to name but a few instructional options.

As an alternative to multiple-choice questions, students can use "mini-white-boards" or slates to write down an answer and hold them up for the teacher to see. This response format is particularly useful in cases where non-standard, but revealing responses are generated by students, but would be less likely to be generated where responses are given to the students. For example, a teacher might ask students to provide an example of a fraction between $\frac{1}{6}$ and $\frac{1}{7}$. In a constructed-response format, many students will write

$$\frac{1}{6\frac{1}{2}}$$

which reveals a sound understanding of the basis of fractions, and provides a useful starting point for a discussion on mathematical notation. At the end of a sequence of instruction, another option is to use “exit tickets”. Every student writes down the answer to the question on an index card, possibly with an explanation of why the answer was selected. The cards are handed to the teacher as students leave the classroom. The teacher can then sort through responses, and use the information to guide planning for the next day’s lesson.

The second and third characteristics of good “decision pull” questions go hand-in-hand. Initially our focus was on the second characteristic listed above—ensuring that the wrong answers were interpretable. This goal was achieved by ensuring that incorrect answer choices were linked to incomplete or primitive conceptions that students have about the content, or other common errors that students make. Although such an approach is sometimes taken by professional writers of multiple-choice items, in our experience this is rarely done both systematically and well.

However, as we thought more about the inferences that teachers would be making on the basis of information derived from these items, it became clear to us that, as well as ensuring that incorrect responses were interpretable, it was as, if not more, important that the *correct* answers be interpretable, i.e. that students who selected the correct answer were doing so because they understood the concept and not that they were applying some form of incorrect reasoning that happened to generate the correct answer. This is well illustrated in the following example.

> *There are two flights per day from Newtown to Oldtown. The first flight leaves Newtown each day at 9:05 and arrives in Oldtown at 10:45. The second flight from Newtown leaves at 2:15. At what time does the second flight arrive in Oldtown? Show your work.*

The problem with this item is that students who fail to realize or remember that there are 60, rather than 100, minutes in an hour can generate a correct solution. Although it is intended as an item on addition and subtraction of times, it is likely that conclusions based on the answers to this item will generate a number of “false-positives”—assumptions that the students have an understanding of the material when they have not. If, as we believe, false positives are more serious than false negatives (assuming that students have not understood something when they do), then it is more important that the *key* is interpretable than it is that the *distractors* are interpretable.

This, of course, is an issue of validity. On the basis of student responses, we are seeking to make inferences about the quality of a student’s cognitive processes, and as Messick (1989) points out, this involves the dual processes of first establishing that the chosen interpretations are supported by empirical and theoretical rationales, and second, establishing that plausible rival inferences are less well supported.

Of course, generating items that satisfy these requirements is more craft than science. Sometimes, very small changes in the item can produce substantial changes in its

functioning. In the item above, just amending the arrival time of the first flight to 10.55 would allow the item to distinguish between students who calculated on the basis of 60 minutes in an hour and those who calculated on the basis of 100 minutes in an hour. In designing items for multiple-choice format, it is also important to note that the interpretability of the key will depend on the quality of the distractors.

Reliability Traded for Utility

It is undisputable that the reliability of a single item is less than the reliability of a testlet, of say ten items, addressing a single concept. And certainly the reliability of a single item would not support high stakes instructional decision-making for an individual. However, the realities of the classroom are such that few teachers make instructional decisions at the level of the individual, but rather at the class level or for groups of students within the class. So for a class of thirty students, with a single question, the teacher has thirty data points on which to base instructional decisions. And given the nature of the constrained choices of a multiple-choice question, students' answers will cluster into a small number of categories, thus further assisting teachers process and interpret the information.

Furthermore in terms of processing ability few teachers would be able to make sense of thirty profiles of performance across ten items in a short enough space of time for the information to have a real-time impact on instructional decision-making. The need is for "just enough" assessment information to be available to in order to meaningfully direct instruction, but not so much that assessment takes up time that would be better spent on instruction. Thus, while little can be concluded with absolute certainty from a single item, such items are likely to be better than the approaches teachers currently use for judging a class's understanding, and the teacher can always ask students to explain their answer if further insight into their thinking is needed.

Furthermore, while the possibility of misclassification exists, the consequences are low. The teacher may re-teach information that students have already learned but such mistakes easily remedied. If, in the course of the additional instruction, it becomes clear that the students already do understand the information, the teacher can move on to the next topic. While the accuracy of a decision based on a single item may be far from perfect, it is better than a decision based on no data at all.

Selecting the Question

In the DIMS project we are not suggesting that all classroom assessment be done in the form of single items. Where summative inferences are required, then the usual array of classroom assessments will be far more appropriate. However, for the purpose of a rapid assessment of student learning, we contend that a single well-chosen question can provide enough information to direct instruction in real-time, provided the item is chosen carefully.

Falmagne, Cosyn, Doignon & Thiéry (2003) show that the efficiency of an assessment of the mastery of a given domain can be increased substantially by making use of the internal structure of the domain. For example, consider the domain defined by the criterion "addition of two numbers under 100". The assessment domain consists of all possible pairs of numbers less than 100. However, the ability to compute 7+3 can be assumed if a student is also able to add 27+53. A subset of the domain is those calculations that require carrying. We can assume that students who can successfully compute 27+55 will also be successful on the easier items. A subset of calculations that require carrying are those that require carrying twice. A student who can successfully compute problems that require carrying twice will have mastery of the entire domain. In the terminology of Wiliam (1993) this subset of the domain provides a *cover* for the domain. If it is the smallest such subset, the subset is a *minimal cover* for the domain. In the item on plane times above, the set of all items that require knowledge of the fact that there are 60, rather than 100, minutes in an hour provide a cover for the entire domain.

Therefore, if all we care about is mastery of the whole domain, then we should select from the minimal cover for the domain. This may sound like an obvious point, but this approach differs radically from the standard psychometric model, where items are sampled at random from the entire domain (Loevinger, 1965 p. 147). Of course, if we wanted to ascribe a degree of mastery to a student, then the traditional psychometric approach would be effective, but even here, knowledge of the structure of the domain can be used. In the same way that stratified random sampling can increase the representativeness of a random sample, provided the stratifying variable is theoretically relevant, then using nested sequences of covers would provide a stronger, and, more importantly, a more interpretable indication of the degree of student mastery of the domain. Connecting the incorrect answers to common student conceptions is a critical part of the development process. A teacher is more able to target instruction if he or she knows more precisely what students are struggling with as opposed to just *that* they are struggling. For example, a criterion might be "identification of the median from a list of up to ten numbers." Students have several conceptions about the median such as "it is not necessary to order the list of numbers" and "an even-numbered list of numbers does not have a median". Starting with common student conceptions and crafting a question such that it is set up to elicit those conceptions helps to identify the *cover* of the domain. Thus a single question that asks students to identify the median from a non-sequential list of eight numbers is likely to provide good cover of the larger domain.

Conclusions

We are not suggesting that a teacher make individual student inferences based on a single item. However, we believe that a decrease in reliability is justified by the accompanying increase in utility provided that single items can be used in situations where the consequences of misclassification is low, and the question is selected in such as a way that it provides good cover of the domain, and where both correct and incorrect answers are interpretable.

References

Falmagne, J.-C., Cosyn, E., Doignon, J.-P., & Thiéry, N. (2003). *The assessment of knowledge in theory and in practice*. Irvine, CA: Institute for Mathematical Behavioral Sciences, Paper 26.

Helmke, A., & Schrader, F. W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education,* **3**(2), 91-98.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: minute-by-minute and day-by-day. *Educational Leadership,* **63**(3), 18-24.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review,* **72**(2), 143-155.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3 ed., pp. 13-103). Washington, DC: American Council on Education/Macmillan.

Popper, K. R. (1976). *Unended quest: an intellectual autobiography*. LaSalle, IL: Open Court.

Wiliam, D., & Thompson, M. (2006). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates.