Specific Mathematics Assessments that Reveal Thinking: an online tool to build teachers' diagnostic competence and support teaching

Kaye Stacey, Vicki Steinle, Beth Price, Eugene Gvozdenko

University of Melbourne

In this chapter we describe the design of an online system for the formative assessment of students' understanding of mathematics and discuss how it develops diagnostic competence and influences teaching. The smart-test system covers many mathematics topics studied by students between about 10 and 16 years of age. It is programmed to provide teachers with an automated diagnosis of their own students' stages of development in specific topics, and to report on an individual's errors and misconceptions, in order to inform teaching. Our claim is that teachers' diagnostic competence increases when they have easy access to information about their own students' thinking. In turn, this can further improve teaching, and hence learning. By drawing together evaluative data from four sources, we highlight aspects of teachers' initial responses to formative assessment, and the effect of using this system on their knowledge for teaching and the subsequent changes to teaching practice. Overall, teachers report that using the smart-tests has improved their knowledge of the thinking of individual students as well as of students in general (i.e. their pedagogical content knowledge), and that they can use this information in several ways to adjust their teaching. Paradoxically, using smart-tests reduces the demand for teachers to have specific knowledge for diagnosis, and at the same time increases this knowledge and so improves their diagnostic competence.

1 Introduction

This chapter discusses the design and use of an online assessment system and presents a partial evaluation of the success of this system. The smart-test system is built on *Specific Mathematics Assessments that Reveal Thinking,* which we abbreviate as smart-tests. These tests are accessed through an intelligent environment (HREF1), created by the authors (Stacey, Price, Steinle, Chick & Gvozdenko, 2009). The goal is to diagnose individual student's understanding of mathematics topics, hence improve the teacher's understanding of student thinking, and thereby assist teachers to target lessons to better meet the needs of their students.

The smart-test system provides teachers with an informative diagnosis of their students' conceptual understanding of many of the topics in the curriculum for students between 10 and 16 years of age. The diagnoses are described in terms of developmental stages and the misconceptions and/or common errors that have been identified for a particular student. These diagnoses are available to a teacher immediately after their students complete the test. As far as possible, the items, the developmental frameworks and the diagnostics are drawn from well-established research findings, and so build in pedagogical content knowledge related to student thinking. In addition, the system provides teachers with explanations of the diagnoses, teaching suggestions for taking students to the next level of understanding and, if appropriate, for dealing with misconceptions and common errors. Although the categories are not well-defined, we find it useful to distinguish between misconceptions (which have an identifiable underlying conceptual base) and common errors (e.g. information that students have not learned, confusions of names, bugs in algorithms) that are procedurally or factually based.

Because smart-tests aim to assist the teacher to plan more effective teaching, a smart-test is focused on one topic and typically takes students less than 10 minutes to complete. To encourage teachers to monitor student progress throughout the teaching of a topic, there are two parallel test versions for each topic. Smart-tests are not intended to be a complete assessment of the topic – for example, there are no lengthy items which require students to write mathematical reasoning. The smart-test system is currently being used regularly by over 400 teachers, and we process approximately 7000 student tests each month. The system can be used by teachers and students anywhere with an internet connection.

The smart-test system is designed to assist teachers with the diagnostic activities of gathering and interpreting data so that they obtain valid knowledge on the achievement of individual students and to provide appropriate teaching suggestions as a basis for action. As will be evident in the sections below, making use of this information involves the whole range of diagnostic competence, including teachers' knowledge (especially their pedagogical content knowledge), beliefs connected to formative assessment and the skills to implement it and to act on the findings. We will demonstrate how the smart-test system paradoxically reduces the demand for teachers' diagnostic competence, whilst at the same time building it.

Section 2 outlines the designers' vision. In creating any educational product, there is a myriad of design decisions, so this paper concentrates on those that are central to providing diagnostic judgments that can support productive action by teachers. Sections.3 and.4 draw together feedback from teachers gathered from several different sources over the life of the project. In section 3 we discuss two themes related to teachers' evolving understanding of the use of formative assessment – what it is really for and how it is best used and discuss how we responded to these issues. In section 4 we report on our progress towards achieving the two fundamental goals of the smart-test system, namely higher achievement for students through targeted teaching, and improved mathematical pedagogical content knowledge for teachers.

1.1 Data sources

This chapter reports experiences of the smart-test creators and feedback from users. The data reported in sections 3 and 4 is a collation from four sources: (1) records of ten focus groups held with teachers at three schools involved in the development of the smart-test system in its first two years (2008 – 2009); (2) online surveys completed by volunteer teachers after they have used a smart-test (2009 – 2014); (3) spontaneous emails that teachers have sent to us on an ad-hoc basis after completing a smart-test (2009 – 2015) and (4) interviews with three mathematics leaders. Two of the interviewees were teachers holding leadership positions in mathematics at their schools (Leader 1 and Leader 2); they were interviewed in 2015. The third interviewee was a Project Officer, employed by an education authority, who assisted teachers using the smart-tests in their own schools as part of a larger professional development learning program. She was interviewed in both 2012 and 2015.

Gaining feedback on the smart-test system is in itself a process of formative evaluation, with the aim of improving all aspects of the system. As is evident from the sources described above, the data sources reflect the long development time for this complex resource. Because of ongoing improvements, the resource to which they responded is somewhat different at each stage with early concerns having now been addressed. The feedback reported in sections 3 and 4 focusses on issues that transcend pragmatic concerns (e.g. difficulty scrolling on long pages, download speeds) and gets to the heart of how teachers might use formative assessment from any source to improve their teaching.

2 The designers' vision for the smart-test system

> "*...because learning is unpredictable, assessment is necessary to make adaptive adjustments to instruction.*" (Wiliam, 2011, p. 13)

The initial concept of the smart-test system arose from our observation of the power of formative assessment and our observation of the difficulty of diagnosing students' thinking quickly and efficiently. This is well supported by others (e.g. Wiliam 2007, 2011). In previous research projects, we saw how teaching about decimal numeration could be transformed by giving teachers information about the ways in which each of their students thought about decimal notation and by helping teachers understand the common misconceptions (Helme & Stacey, 2000).

However, we also noted that it is time-consuming for teachers to diagnose individual student thinking using research level tests (Steinle, 2004) or simplified versions (Steinle, Stacey and Chambers, 2006). Diagnosis from written tests is usually complicated because special scoring instructions must be followed to identify the tell-tale patterns in students' responses that indicate a misconception. Teachers expect to mark students' work as correct or incorrect, and to make total scores or sub-scores, but it is beyond their expectations that they should undertake further processing of data, such as examining patterns of responses rather than just observing direct errors in a systematic way. The solution to this dilemma was to use online assessment, with computer programming identifying the patterns of responses across multiple items that reveal thinking. Hence the smart-test system began.

The need for technological help in formative assessment has been noted by others. For example, Pellegrino and Quellmalz (2010) wrote:

> *"No individual, whether a classroom teacher or other user of assessment data, could realistically be expected to handle the information flow, analysis demands, and decision-making burdens involved without technological support. Thus, technology removes some of the constraints that previously made high-quality formative assessment difficult or impractical for a classroom teacher." (p. 130)*

We planned that this system would be easy and efficient for teachers to use and that it would supply information that is concise enough to be readily useable by teachers, sufficiently valid and deep enough to make a real difference to lesson content, and linked to targeted teaching resources. Figure 1 shows how we expected teachers to interact with the system and the two predicted outcomes: higher achievement for students through targeted teaching, and improved mathematical pedagogical content knowledge for teachers. We expected that this aspect of knowledge for teaching would improve as teachers become familiar with the developmental stages and possible misconceptions in a particular topic, especially as they see how these apply to their own students.

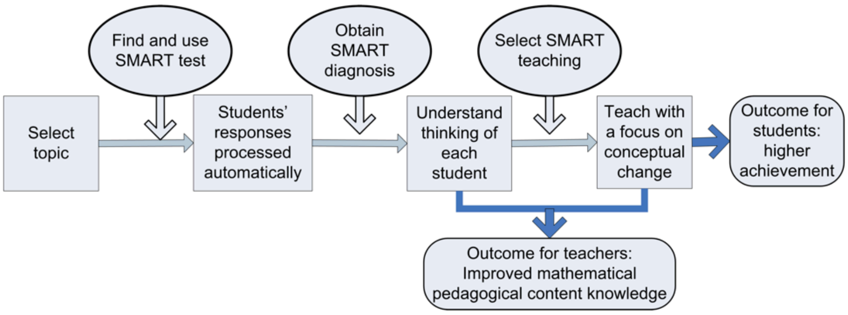

Find and use SMART test
Obtain SMART diagnosis
Select SMART teaching
Select topic
Students' responses processed automatically
Understand thinking of each student
Teach with a focus on conceptual change
Outcome for students: higher achievement
Outcome for teachers: Improved mathematical pedagogical content knowledge

Figure 1. Using the smart-test system and its predicted outcomes

The smart-test system embeds research in mathematics education into artefacts that are intended to be easy for practitioners to use, creating what Pea (1993) calls 'distributed intelligence' in tools for teaching. When planning the teaching of a new topic, the diagnostics from the system provide teachers with knowledge of the mathematical thinking of their current students. It is intended that, simultaneously, teachers will also learn about the mathematical knowledge of students more generally and hence will be better able to teach effectively in the future.

There are considerable benefits if a teacher is able to conduct interviews with students on their mathematical understanding. Indeed, interviewing all students to establish their stages of development has been a central feature of the highly effective early numeracy programs in Australia, such as Count Me In Too (Stewart, Wright & Gould, 1998). There are further examples in the chapter by Clarke et al. in this book. However, interviewing is a resource-intensive option, for which schools need to make very special arrangements. We make no claim that smart-test information is always completely accurate, but neither is any other method (although this is sometimes not recognized!) and teachers can choose to talk to those few students with puzzling smart-tests results. The intention is that the smart-test system should provide teachers with sufficiently valid information to influence the teaching of topics about to be taught, in a timely and simple manner.

2.1 Items that provide a window into student thinking

> *"Items that reveal unintended conceptions – in other words that provide a 'window into thinking' – are not easy to generate, but they are crucially important to improve the quality of students' mathematical learning."*
> (Wiliam, 2007, p. 1069)

Smart-tests are built on exactly the sorts of items that Wiliam refers to in the quote above: sets of items that together provide a window into student thinking. Over the several decades of mathematics education research into students' thinking, a rich bank of items has been established which smart-tests make more accessible to classroom practice. However, items usually need to be modified for use in computer-based assessment because a computer is still limited in its processing of free response items (Stacey & Wiliam, 2013). Hence smart-tests often include selected response items with alternatives based on research evidence.

In addition to the multiple choice format, there are now other selected response formats that can be readily computer marked. Sliders provide a very flexible interactive format. Students can place numbers on number lines as in Figure 2, and show estimates of quantities like percentages and angles. Drag-drop items allow students a different type of participation, similar to the way they might participate in an interview. For example, students can arrange 'cards' showing various fractions and decimals in ascending order; they can place a card showing the position of the translated image of an object; and they can drag cards showing representations of the fraction two thirds into one pile, and other cards into another. Figure 3 shows a student's incorrect pattern of responses to an item about sorting angles. Using this expanded range of computer-assessable formats has made the tests more interactive.

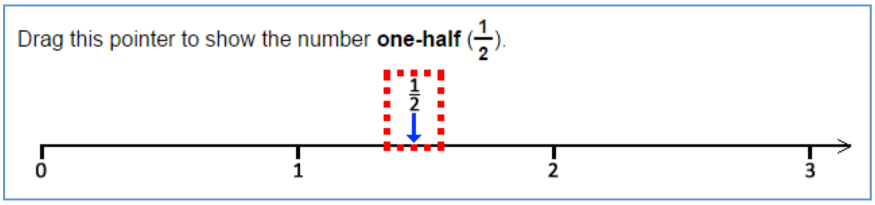
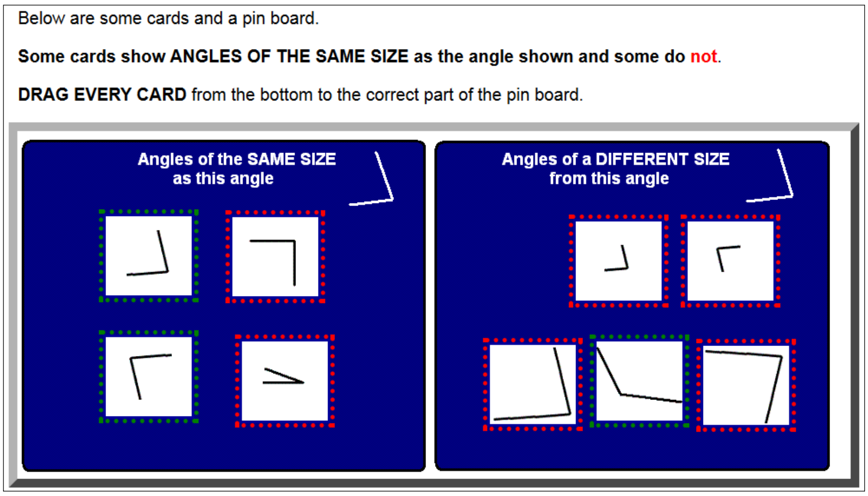

Drag this pointer to show the number **one-half** ($\frac{1}{2}$).

Figure 2. A student's incorrect response to a slider item


Below are some cards and a pin board.

Some cards show ANGLES OF THE SAME SIZE as the angle shown and some do not.

DRAG EVERY CARD from the bottom to the correct part of the pin board.

Angles of the SAME SIZE as this angle
Angles of a DIFFERENT SIZE from this angle

Figure 3. A student's incorrect response to understanding angle size item

As well as using items with strong credentials from the research literature, sets of items for smart-tests must systematically vary the features which are known to make a difference to item success rates. This enables students' difficulties to be pinpointed. It also gives guidance to teachers about the range of items to include in instruction and what constitutes robust understanding. Using only prototypical item types in teaching is known to encourage misconceptions and limited understanding. For example, when considering angles of the same size, factors such as the ray length shown and the orientation of the angles affect success rates. The drag and drop item described earlier identifies whether students see the angle (the amount of turn) by systematically varying distracting factors. Figure 3 shows the display seen by a teacher for one of their own student's choices of angles of the 'same size'. The borders that can be seen in this figure are to assist teachers when they look at test submissions; green is correct and red is incorrect. Note that this student, like many others, has incorrectly used the visible ray length to classify the angles. This student has done this consistently; others may nearly always do it. Teachers are provided with a brief summary of findings for each student, and have the option of accessing student screens to see the behavior in action.

2.2 Diagnosis from patterns of responses, not just score

Many mathematics tests base their assessment only on accuracy, either total test score or scores on subsections. The power of the smart-test system is that it diagnoses student thinking based on the actual responses. Responses (more than accuracy) are used because 'wrong in which way' is more revealing than just 'wrong'. For example, the student whose work is shown in Figure 3 might be considered to have a score of 3 out of 9, but the pattern of wrong answers actually shows *why* they were wrong, not just that they *were* wrong. As far as possible, smart-tests report on the reasons for errors, not just the number of items correct.

Steinle, Gvozdenko, Price, Stacey and Pierce (2009) indicate how response patterns can be used to diagnose student misconceptions in algebra in the test named *Values for letters*, which draws on extensive research in algebra such as Küchemann (1981) and Fujii (2003). Those who sign-up on the smart-test website can access the 2012 version of this test which replaced the 2009 version. One set of items in the test *Values for letters* describes the scenario that 'some students' were asked to find the values of letters in several equations. In the first item, the students taking the test are asked to indicate whether the solutions to $x + x + x = 12$ given by the fictional students are right or wrong. For example, one fictional student has answered '$x = 2$ and $x = 5$ and $x = 5$' (which is incorrect but accepted by students who see the letter as simply a placeholder for any number) and another has answered '$x = 4$' (which is actually correct, but will be rejected by students who want a value for each of the three occurrences of x). In the second item, the equation is $x + y = 16$ and the solutions given by the fictional students include '$x = 7$ and $y = 9$' which is correct, although rejected by the few students who think y should be one more than x because y is one letter after x in the alphabet (Stacey & MacGregor, 1997) and '$x = 8$ and $y = 8$' (also correct but rejected by students who believe that different letters must stand for different numbers). Based on the pattern of their responses, students are allocated to one of four developmental stages (see Figure 4) for interpreting the letters in equations and are flagged when they have certain misconceptions. Smart-test items are often, as in this case, drawn from research literature, but using the tests also provides data on the prevalence of reported misconceptions in our population. In a few cases previously unreported misconceptions have been revealed.

The only situation where we use the score (total number of responses correct) to provide information to teachers is for an additional feature introduced in response to feedback from teachers about anomalous results. Occasionally a student fails to clear the hurdles for the early stages, yet answers more difficult items correctly. Students with a high score on a test but a low developmental stage are therefore flagged so that teachers can investigate individually whether the students do have a fundamental misunderstanding or gap in their knowledge, perhaps masked by procedural expertise, or have just made some careless errors or omissions. We have observed examples of both situations. As an extreme example, a student who did not answer any of the 'easy' items for some extraneous reason (strange things sometimes happen in classrooms!) would be flagged as not having met criteria for the first stages, even though they may have answered the most advanced items correctly. A more common example might be a student who is expert at the addition algorithm for fractions and uses it to solve some complex questions (albeit in a complicated way), but who cannot answer apparently more elementary questions about the meaning of fractions.

2.3 Reporting developmental stages and misconceptions

> "[Evidence generated to support learning needs to be] *more than information about the presence of a gap between current and desired performance. The evidence must also provide information about what kinds of instructional activities are likely to result in improving performance.*" (Wiliam, 2011, p. 11)

A major design decision has been how to present results to teachers. We wanted to use computing power to move away from using behavioural item-by-item descriptions (i.e. saying what items student get correct) to look at broader stages of conceptual development that give teachers more insight into student thinking. We also wanted to help teachers understand how students perceive mathematical ideas. The approach that we selected is to describe learning in terms of topic specific stages along a learning hierarchy. Our website calls these 'developmental stages'. We decided to report on each student's stage in the specific topic and also flag if they exhibited any misconceptions or common errors (Stacey, Price, Steinle, 2012).

A learning hierarchy is created by considering some combination of the following: postulating a complexity order based on logical analysis; using teaching experience; using prior research and analysing empirical data. Stages in a learning hierarchy are confirmed by data if several conditions are met. We require items with similar mathematical characteristics to have similar success rates, and to be completed successfully by the same students. If these conditions are not met, the items need further investigation. When a learning hierarchy exists, knowledge at one stage is pre-requisite for achieving tasks at a higher stage. This means that students unable to complete items designed to test lower stages will be less likely than other students to successfully complete items designed to test higher stages. There are many complexities in this simplified story: for example, some of our learning hierarchies have branches. A detailed example is given by Stacey, Price and Steinle (2012).

Figure 4 shows the four developmental stages and the misconceptions and common errors created for the 2012 *Values for letters* smart-test described above. This is a later version of the test and developmental stages than reported by Steinle et al. (2009) because analysis of more student data caused us to add more items, and so improve the reliability and range of diagnosis. More items improve the reliability of diagnosis because there is more capacity to discount the influence of careless (i.e. unsystematic) errors and hence more capacity to be sure that patterns in responses are a true reflection of students' thinking. The new items explore alphabetic misconceptions and they check that students have the basic knowledge of substitution to complete the items meaningfully. The developmental stages go from early use of letters as a code for numbers, to basic understanding of letters as a 'place holder' for a number, and on to refined understanding. Three of the misconceptions relate to alphabetical interpretations of algebraic letters that can linger to influence students' thinking at various levels of competence, and the fourth is an interesting error that reveals lingering uncertainty about letters as placeholder (students rejecting '$x = 4$' in the item above but accepting '$x = 4, x = 4, x = 4$').

These stages are empirically confirmed, although this does not mean that every student "goes through" each of these stages. Well taught students may, for example, very quickly jump from Stage 1 (a basic idea of a letter as a place holder for a number) to Stage 4. In fact, this test is unusual in that the developmental stages might also be seen as identifying misconceptions and a somewhat arbitrary (but empirically confirmed) decision has been made to put one sequence of misconceptions into the stages and flag others separately. This is due to the very specific nature of this test. A broader test (see for example Stacey et al. 2012) uses larger steps of new knowledge for the basis of the stages.

There are many design decisions related to the presentation of results so that their usefulness to teachers is maximised. In order for the smart-test system to improve teachers' pedagogical content knowledge, information needs to be accurate, complete and well-researched, but busy teachers are unlikely to spend considerable time reading a large amount of text. Space precludes discussion of most of these design issues. Relevant to this paper is the major issue of finding the appropriate level of detail and of technical language to describe stages. To this end, we have prominent brief versions (as in Figure 4) hyperlinked to detailed explanations and examples. Similar considerations apply to the teaching advice which is given for students at each stage.

Developmental stages for smart-test *Values for letters*

Stage 1: The students know that letters can stand for numbers, and are able to correctly substitute into very simple algebraic expressions, but they believe that the values that letters can take are in some way related to their place in the alphabet.

Stage 2: These students interpret an algebraic letter only as a place holder for a number in a number sentence, so they allow one letter to have several values in one expression.

Stage 3: These students appreciate that each time a particular letter is used in an equation it stands for the same number, but they over-generalise to 'different letters must be different numbers'.

Stage 4: These students know that in one algebra question, a letter must stand for only one number and that different letters can stand for the same number.

Misconceptions and common errors for *Values for letters*

A: Students often give a letter a value related to its place in the alphabet, such as b = 2.

C: Students believe that the values of consecutive letters must be consecutive numbers.

O: Students believe that if one letter is before another in the alphabet, its value must be smaller.

R: When the same letter is used more than once in an expression, students wish to state this value separately for each occurrence.

Figure 4: *Values for letters 2012* – summary of stages and misconceptions

3 Helping teachers transform diagnostic information into formative assessment

The sections above have presented smart-tests from the point of view of the designers. As well as analyzing student data to improve the items and diagnoses, we have sought feedback (see section 1.1) from teachers throughout the development of the system. As a result, many additional features have been included, a few of which have been mentioned above, and we have also developed advice for teachers and school leaders to use smart-tests effectively. In essence, this section demonstrates that, while the smart-test system takes over some aspects of diagnostic competence, other aspects of diagnostic competence are needed to use the tool effectively. Because smart-tests are different from the tests that teachers normally set and students take, teachers are likely to initially experience some disequilibrium. Here we explore two frequently raised issues which are related to this disequilibrium: teachers' assumptions about the purpose of the assessment (summative or formative); and their request for the system to provide feedback directly to students.

3.1 Appreciating formative assessment

We found that the concept of formative assessment – assessment that directly feeds into lesson planning – was not understood by all teachers, even in the three schools which had volunteered to trial our first diagnostic tests. During the first two years of development, we visited the participating teachers and schools several times each year to get feedback from teachers and conducted formal focus groups. One of the questions regularly raised by the teachers was how to use the smart-test system information in their bi-annual formal reports to parents. These group discussions showed us that teachers primarily wanted summative rather than formative assessment. Teachers have many demands made upon their time and some of the teachers in the participating schools were hoping to use the smart-test system directly in the time-consuming task of writing reports. Some teachers also expected a measurement of the overall level that students had reached in mathematics against the published state standards both for reporting to parents and for accountability to the local department of education. At that time, an increase of one level in the published state standards indicated 6 months of average mathematical growth.

An early issue arising was therefore whether we should adapt our assessment system to meet teachers' desire for summative assessment. We rejected this proposal for various reasons, and retained the focus on diagnosis to inform immediate teaching. For example, there is no easy marriage between global reporting levels and our stages – our developmental stages remain 'local', referring only to the topic. Stage 1 in one topic has no particular relation to Stage 1 in any other. Stages are also not related to the global achievement levels of any official curriculum documents, even though at first glance this might make summative reporting much easier for teachers. One reason for this disparity in stages is their different 'grain size'. Sometimes, diagnostic developmental stages are very fine grained – we intend that students can move through very quickly with appropriate instruction, possibly in a couple of lessons. The stages for the test *Values for letters* in Figure 4 are like this. On the other hand, we know that for some tests where complex concepts are involved, students on average take some years to move through the stages. For example, one of our smart-tests, described in Baratta, Price, Stacey, Steinle and Gvozdenko (2010), looks at percentage problems with different quantities unknown and numbers of varying complexity, and in this case the stages take some years for many students to master. We saw that many teachers took quite some time to broaden their understanding of useful assessment to include formative as well as summative purposes and to appreciate that data on students could be used in the short-term to match instruction to students' needs. One very successful approach to building up an appreciation of formative assessment (confirmed by internet traffic of use and re-use of smart-tests) was including smart-tests in the professional development organized by the Project Officer. She worked with about 50 schools in 2009 to 2015, and visited teachers to assist with implementation. In the programs, teachers gathered data from their own classes through the smart-test system and used the diagnoses in their lesson planning over an extended period. Results from particular smart-tests were also used in the face-to-face sessions to show teachers the relevance to their own teaching.

The Project Officer, when first interviewed in 2012, articulated two common views that she had encountered during early professional development sessions (see section 1.1) which she felt hindered teachers' uptake of formative assessment. The first view was that teachers already know the students in their classes and so can accommodate their needs without any additional specific information. These teachers observe the general mathematical achievement of their students and often plan lessons or goals for broad groupings of high, middle and low achieving students. They do not, however, consider the particular understanding that each student or groups of students have, in different mathematical topics. Instead they tend to expect students to master simple, medium or complex aspects of a topic. The second view is that in every class there is a variety of students, so the whole scope of each topic needs to be covered with the whole class. Therefore there is no need to identify individual student understanding as every student will do every aspect of the topic at the expected level of difficulty. The Project Officer commented in 2012 that both of these views were prevalent among teachers and enabled them to teach in the same way each year with little or no differentiation for differences between classes or individuals.

In the interview three years later, she noted that there has been movement in the use of assessment in schools in the education system in which she conducts the professional development programs, and some shift in the views expressed by teachers. The teachers have been encouraged by the local education authorities to change their view of what 'knowing your students' actually means. She reported that she had not heard teachers saying that the whole class should go through the whole scope of the topic for several years. There is also some evidence from another study (Quenette, 2014) that the views that some teachers hold about their students being either 'good' or 'bad' at mathematics are ameliorated when smart-tests help them appreciate that students who are not making progress may be held up by a gap in knowledge or a misconception, and that even capable students can have misconceptions in some areas.

During the 2012 interview the Project Officer reported that the turning point for the appreciation of formative assessment for some of her professional development participants came when they realised that formative assessment could sometimes save teaching time, by increasing efficiency of learning. She also noted the importance of the first practical use of a smart-test being undertaken within the supportive professional development environment. Examining the results helped teachers realise that their assumptions about student understanding and competence were not always accurate, and that some students who are procedurally competent have misconceptions. Some teachers realised for the first time that they might accidentally write a test on which students with misconceptions could score 100% if they do not use an appropriately wide range of items types (including non-prototypical) and items which probe thinking. Procedural fluency often masks lack of understanding and the latter can hinder long term progress.

For teachers not in a professional development program, the smart-test system provides some information about the use of formative assessment when teachers sign up to use the system. The information stresses its special and different character and gives examples of how it can be used.

3.2 The provision of feedback directly to students

Another issue that many teachers raised early in the project is whether the automated diagnosis should be delivered to teachers only or also directly to students. There are two drivers of this request. Firstly, most computer games or quizzes that students use provide immediate feedback (usually right/wrong), so students expected this from online tests. Secondly, many teachers are aware that good quality feedback, presented to students soon after the completion of a task, can lead to increases in learning. In fact, involving students in the results of assessment is often cited as a hallmark of good formative assessment (see, for example, Sadler, 1989).

After consideration, we have maintained our position to provide information only to teachers. We want teachers at the centre of the diagnostic process, because we believe that substantial teacher input is required to overcome most of the conceptual obstacles identified. Furthermore, the detailed topic-specific diagnoses are written for adults, and some effort, background and technical language is required to understand them. Student feedback would need to be written separately and at a variety of reading levels. Leader 1 was interviewed about a range of implementation issues in her school. In the interview she described how teachers adapt the smart-test feedback they receive to describe in age-appropriate terms for students what the student has shown he or she can do and also the stage toward to which he or she will be working.

Another issue in potentially providing diagnoses direct to students was the disagreement among teachers on the nature of the feedback that might go directly to students. For example one group of teachers warned about negative consequences of students receiving feedback which indicated low performance and hence they recommended only good performances to be reported to students.

We have now resolved the student feedback dilemma by advising teachers to explain the purpose of formative assessment to students before the test, so students know that the information gathered will be used to their advantage, whether as individual feedback or to adapt class lessons. Feedback from teachers, such as Teacher A in Figure 5, indicates that this is a successful strategy. The observations by the Project Officer (see Figure 5) confirm this general impression. As a separate project, we are currently developing a modified system which provides diagnostic information direct to test-takers who are pre-service teachers.

Teacher A: “I just explain to the kids what it shows, and that it’s showing me how to teach […] better. ‘It’s not about something that you’re going to get tests back. It’s just a tool that I’m using to see what you guys know, so I can teach you better.’ They have had no issues with that. And the parents that I’ve spoken to on parent teacher night a couple of times think it is fantastic.”

Project Officer: “Teachers explained the purpose of the assessment to students and … the students became relaxed when the teacher followed through with the intent.”

Figure 5. Comments related to providing assessment results to students.

4 Building diagnostic competence and improving teaching

Our aim for the smart-tests is to improve student learning and also to improve teachers’ knowledge for teaching (especially their pedagogical content knowledge). In this section, we report on teachers’ views about the effect of using the smart-test system on their knowledge for teaching and whether (as well as how) it has changed their practice. Achieving these two goals will really make building the smart-test system worthwhile.

4.1 Effect on knowledge for teaching

As is shown in Figure 1, one of the aims of the smart-test system is to increase teachers’ mathematical pedagogical content knowledge. We hypothesised that putting data on their own students’ thinking into teachers’ hands would make research results come alive for teachers, and hence build their capacity to teach both current and future students. In this sense, the smart-test system is trying to take research results out of the library and put them into the hands and minds of teachers.

In his 2015 interview, Leader 2 expressed the opinion that he found smart-test diagnostic information useful for teacher learning:

> *“It is difficult to find time to lift the mathematics content knowledge of the teachers. …. This is an ideal way of me being able to introduce a little bit of professional development informally... I feel easier because I know that the teachers are getting some professional development.”*

Some data on whether teachers feel their knowledge base has improved comes from the voluntary survey, completed by teachers after they accessed students' results from a test. The survey included multiple choice items, with space for optional comments on each item. Table 1 provides the frequency of survey responses to the multiple choice question: *As a result of using this quiz, have you learned something useful for you as a teacher?* The results show that nearly all respondents reported that they learned something useful and nearly half chose "very valuable learning".

Table 1. Online survey responses to increasing pedagogical content knowledge

Options provided	Frequency	Percent *
YES, very valuable learning	115	47%
YES, useful learning	117	47%
NO	15	6%
blank	16	-
Total	263	

*Percent of 247 non-blank responses

The associated comments provide further evidence that teachers perceived that using the smart-test system has led to improvement in their knowledge for teaching. Sample comments are provided in Figure 6. Both teachers B and C reported their own increased confidence in understanding how students think, whilst teachers D, E and F added successful new teaching strategies to their repertoires from the teaching advice provided by the system.

We acknowledge that self-reporting has limitations as a method of data collection, both because of the volunteer sample and in the opinions expressed, so we are cautious in the use of the data obtained. We expect those with strong opinions for or against the smart-tests to be over-represented. Since there was no pressure on teachers to make positive comments about smart-tests (and indeed our questions encouraged specific suggestions for system improvement) we expected that the direction of the comments would indicate the general feeling but that the strength of the opinions offered would be stronger than that of the general teacher population.

Teacher B: "It certainly has encouraged a dialogue between the student and the teacher, and looking at specific things because you as a teacher feel more confident about what you're talking about, because you've got all that information there. The smart-test directs you about where to go. And also you can speak to that student about that particular misconception. It works quite well."

Teacher C: "Well worth doing. Made me feel like an 'expert' teacher instead of just an experienced teacher."

Teacher D: "[I] used a table structure similar to dual number line to help students with showing and organising information contained in problems and to find what 1 part represents [and also] to emphasize the use of multiplication/division."

Teacher E: "I WILL use more materials and a lot more justification from the students." (emphasis used by teacher)

Teacher F: "I read the referenced research paper, which was informative and useful. The teaching suggestions were really practical, and were suitable to have a go at straight away. I used paper strips and pieces to fold and colour to estimate percentages."

Leader 1: "I think teachers are now (since they have been using smart-test data for planning) more confident in … identifying where the students are at."

Figure 6. Comments related to improvement in knowledge for teaching

The very high proportion of "yes" responses in Table 1 is likely to be due to the fact that many of the teachers were using a particular smart-test for the first time when they completed the surveys. We expect that, on subsequent use of the same test, teachers will be more familiar with the developmental stages, and so they are unlikely to report valuable 'new' learning, except through the results of individual students. In fact, we intend that they will come to observe the developmental stages and misconceptions in their normal interactions with students. In this way, we hope the smart-tests may become redundant, as teachers modify their teaching to reduce the likelihood of misconceptions, help students to develop strong conceptual understanding, and have at their fingertips items which reveal understanding in the specific topic. For example, after knowing about the phenomena revealed by the *Values for letters* test described above, teachers can quite easily address students' false assumptions in their teaching and take care to look for them in students' work. If the test is no longer required because of increased teacher understanding of students' thinking, then that is itself a success.

4.2 Effect on teaching practice

An assessment is only formative if it results in a change in the opportunity for a student to learn. Table 2 provides the frequency of responses to the voluntary survey multiple choice question: *Did you adjust your teaching plan as a result of the diagnostic information?* Of the 220 responses to this question, 70% indicated that they did adjust their teaching. Of course, adjusting is not always required. One of the teachers who did not adjust their teaching commented: "I didn't adjust my teaching plan as such, because the results supported what I expected, but confirmation was valuable”.

Table 2. Online survey responses to teachers adjusting teaching plan

Options provided	Frequency	Percent *
YES	154	70%
NO	66	30%
blank	41	-
Total	261	

*Percent of 220 non-blank responses

The follow-up question to teachers was: *If YES: In what way did you change your teaching plan?* There were many different types of responses. Two very frequent themes are illustrated in Figures 7 and 8. The first theme, illustrated by comments in Figure 7, was that teachers used smart-test information to target teaching to specific groups or individuals, especially for overcoming misconceptions or revising basic knowledge.

Teacher G: “I often overlook and brush off students’ misconceptions without considering the difficulty that students face. With this assessment tool, I am able to analyse my students better individually, and correct their misconceptions on a particular topic.”

Teacher H: “I have put the students into groups and will give them activities to focus on and correct their misconceptions. I will be looking carefully at the [suggested resources].”

Teacher I: “Very useful as a pre-test on reading scales. I found out exactly where each student was at and that enabled me to target my teaching into the areas where it was most needed, while giving extension work to the students who had already gained a good understanding of the topic. Now I am going to retest them using another form of the test to see how effective my teaching has been.”

Teacher J: “I had assumed that at year 10 my students would have a basic understanding of the idea of percentages - many of them didn't! Instead of going straight into calculating percentages of quantities and calculating whole quantities given a percentage, and then on to financial arithmetic (simple interest), I went back to basics with the students who needed it, and others who could cope with this were assigned the original tasks I had planned.”

Figure 7. Comments related to changes at the level of individuals or groups

The second theme concerned changes to the starting point for a unit of work for the whole class. Teachers K, L, M and N (see Figure 8), are examples of the many who commented that they started their teaching of the tested topic at a higher level than they had earlier planned. We had expected that many teachers would be alerted to students' inadequate preparation for learning a topic, and so have to start their teaching at a lower level than expected as described by Teacher O, but the comments showed that the opposite situation also commonly occurred.

Teacher K: "I adapted the simpler task that we were going to approach in class with something that reflected the students' greater level of understanding."

Teacher L: "I used the smart-test 'Understanding angle' with my year 7 class. In my teaching I adopted an approach that best addressed the needs of the students based upon the diagnostic test. I was able to avoid certain areas that were well understood and concentrate on areas that were not."

Teacher M: "I looked at the course outline. As many of my students were very strong in perimeter, we focused more on area and volume."

Teacher N: "The other end of the spectrum is that I've been more confident in moving kids, not making them go over things. I can see 'alright, this child has a really good understanding of fractions'. I'm not going to … make him (or her) repeat all of those skills so I feel more confident in moving them to something else."

Teacher O: "When our Year 7 students did the fractions smart-tests, we were surprised to find many students were at Stage 0. All these years we've always presumed that they were at a particular level but obviously that's not happening, and so that's changed our curriculum, the way we think about teaching fractions."

Figure 8. Comments related to changing starting points for teaching

In the 2012 interview, the Project Officer reported her observation that some teachers had become more centred on the individuals in the class. More often, teachers now planned in more detail for their particular class instead of using the methods that they always used for a particular topic.

Since the smart-test system is used, in the main, by volunteer teachers, and only some of these volunteers spend the time to fill in the survey in a detailed way, it might be expected that general feedback on the system is biased towards the positive. (Negative feedback tends to focus on small technical issues.) Even with this caution, it is good to know that teachers like P, Q, and R in Figure 9 find the system very helpful.

Teacher P: "This quiz is a genuinely useful tool to assist in the differentiation of the curriculum. It is efficient and informative."

Teacher Q: "Excellent formative assessment tool which allowed me as a coach to discuss the various misconceptions and student thinking within a year 8 class.

It provided teachers with real data that allowed them to address the misconceptions through their teaching.”

Teacher R: “I use the smart-tests as a part of my diagnostic 'toolbox'. They are clear, easy to access and give a quick snapshot of where my students' prior knowledge is developed or underdeveloped. This information influences the activities I implement in class, ensuring that the students are being challenged in Mathematics.”

Figure 9. Comments on usefulness of smart-tests for formative assessment

The comments above show that the smart-tests have had an impact on the teaching of some individual teachers. However, we have observed that for some teachers, data from smart-tests seems to be better used collaboratively to inform changes to teaching, at least at the beginning. The professional development initiatives of the Project Officer, as outlined above, provide evidence of this. A similar process was implemented in their own schools by Leaders 1 and 2 and is happening in a small number of other schools. Collaborative work seems particularly important in primary schools where the majority of teachers are not used to dealing with large volumes of data. Leader 1 reported that primary teachers at her school appreciated help with sorting students into groups based on the data provided. At her school, the administration supported joint unit planning:

> *“I have a really supportive administration ….and they can definitely see the benefit of it. We do have an hour planning for each year level for maths each week. If we didn’t have that it would be really hard to do this.”*

Leader 2 also said that there had been some ‘creative’ timetabling at his school to enable teachers to analyze smart-test data together and to plan lessons.

Some data has been collected from teachers working with Leader 1 that may give less subjective information about the effect of smart-test use on the knowledge of teachers. Prior to using smart-tests, these teachers were interviewed about what sort of student difficulties they would expect when teaching various topics from the mathematics curriculum. They were asked how they would explain some key concepts. It is intended that a similar interview be given after a year of smart-test use. The project is not yet complete.

One important indirect measure of the usefulness of smart-tests is their rate of use. Each year from 2008 to the end of 2015 the usage figures have increased. We also track use and re-use of individual tests. When funding allows, we are planning to investigate whether student performance data has improved in schools where smart-tests are routinely used.

A design project, such as the development of the smart-test system, depends on user feedback, especially for polishing the myriad of features that any system has. The data above has principally been collected by us to improve the system, and to ensure that we are offering some teachers a product which they find valuable. It has been successful for this purpose. Other data has also been collected within the educational systems to ensure that funding the system is a good use of their resources. Research to scientifically investigate whether the system does improve student learning outcomes in general requires a different methodology, including careful examination of how the tests are used within the school, both with teachers and students.

5 Conclusion

The intention of the smart-test system is to take the results of research about students' understanding of particular mathematics topics and to embed it into an intelligent system: a tool holding distributed intelligence which amplifies what teachers can do. This paper has reported the views of the early and current users of the system. In general, the surveys report positively on the tests individually and on the system as a whole. However, the wider experience of creating the smart-test system shows that formative assessment is only beginning to be part of the culture of all schools in our region. Some schools are certainly ready for it, and indeed are now actively using this as a standard part of their planning and teaching. Making formative assessment easier through online tools should promote its use, but it also seems important to have professional development showing its advantages and distinctive features, and to provide teachers with advice on implementation. Finally, data from the surveys provides considerable evidence of a self-reported increase in teachers' pedagogical content knowledge and that teachers are using the information in their subsequent lessons.

The smart-test system is an intelligent tool, which is designed to reduce the work that diagnostic activities require of teachers. By providing carefully designed items, many based on research literature, and automating the diagnosis rubrics, the pedagogical content knowledge required by teachers is also reduced. However, as demonstrated in the chapter, diagnostic competence involves more than this – including understanding the purpose of formative assessment, and having the skills to implement it. In summary, diagnostic competence is still required to use smart-tests well whilst in the other direction, the evidence presented shows that using the smart-tests can itself increase diagnostic competence.

Acknowledgements

An early version of this paper (Steinle & Stacey, 2012) was presented at the 12th International Congress on Mathematical Education. The smart-test system was developed at the University of Melbourne by Kaye Stacey, Vicki Steinle, Beth Price and Eugene Gvozdenko, with initial funding from the Australian Research Council Linkage Projects Program LP0882176 in partnership with the Department of Education and Early Childhood Development, Victoria, then further contracts with DEECD and the Catholic Education Office (Melbourne). We thank Ms Valerie Everist, Ms Sara McKee and Mr Terence Griffin for assisting teachers with using this system and for participating in interviews, and all the teachers and students who have contributed to this project to date and continue to do so.

References

Baratta, W., Price, E., Stacey, K., Steinle, V., Gvozdenko, E. (2010) Percentages: The effect of problem structure, number complexity and calculation format. In L. Sparrow, B. Kissane, C. Hurst (Eds.) *Proceedings of 33rd annual conference of the Mathematics Education Research Group of Australia*. (pp. 61 – 68) Fremantle: MERGA. http://www.merga.net.au/documents/MERGA33_BarattaEtAl.pdf

Fujii, T. (2003). Probing students' understanding of variables through cognitive conflict problems: Is the concept of a variable so difficult for students to understand? In N. Pateman, G. Dougherty, J. Zilliox (Eds.), *Proceedings of the 27th annual conference of the International Group for the Psychology of Mathematics Education*, 1, 49-65. Hawaii: PME.

Helme, S. & Stacey, K. (2000). Can minimal support for teachers make a difference to students' understanding of decimals? *Mathematics Teacher Education and Development*, 2, 105 – 120., http://www.smartvic.com/smart/research/index.htm, Accessed 7 Jan 2015.

Küchemann, D. (1981). Algebra. In K. M. Hart, M. L. Brown, D. E. Küchemann, D. Kerslake, G. Ruddock, & M. McCartney (Eds.), *Children's Understanding of Mathematics:11-16* (pp. 102-119). Oxford, U.K.: John Murray.

Pea, R. (1987). Practices of distributed intelligence and designs for education. In G. Salomon (Ed.), Distributed cognitions: Psychological and educational considerations, 47-87. Cambridge, MA: Cambridge University Press.

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the Integration of Technology and Assessment. *Journal of Research on Technology in Education*, *43*(2), 119-134.

Quenette, J. (2014). Diagnostic testing and changes to teaching practice in Year 9 mathematics classes. Unpublished Master of Education thesis, University of Melbourne. Retrieved 3 September 2015 from https://minerva-access.unimelb.edu.au/handle/11343/43027

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119-144.

Stacey, K., & MacGregor, M. (1997). Students' understanding of algebraic notation: 11–15. Educational Studies in Mathematics, 33, 1–19.

Stacey, K., Price, B. & Steinle, V. (2012). Identifying stages in a learning hierarchy for use in formative assessment – the example of line graphs. In J. Dindyal, L. P. Cheng, & S. F. Ng (Eds) *Mathematics Education: Expanding Horizons* (Proceedings of the 35th Annual Conference of Mathematics Education Group of Australasia), 393 - 400. Adelaide: MERGA.

Stacey, K., Price, B., Steinle, V., Chick, H., Gvozdenko, E. (2009). SMART Assessment for Learning. Paper presented at Conference of the International Society for Design and Development in Education, Cairns, Australia. September 28 – October 1, 2009. http://www.isdde.org/isdde/cairns/pdf/papers/isdde09_stacey.pdf

Stacey, K. & Wiliam, D. (2013). Technology and Assessment in Mathematics. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick, and F. Leung (Eds) *Third International Handbook of Mathematics Education*, (pp. 721 - 752). Springer.

Steinle, V. (2004). Changes with Age in Students' Misconceptions of Decimal Numbers. Unpublished doctoral thesis, University of Melbourne. Retrieved 3 September 2015 from https://minerva-access.unimelb.edu.au/handle/11343/39024

Steinle, V., Gvozdenko, E., Price, B., Stacey, K., Pierce, R. (2009). Investigating students' numerical misconceptions in algebra. In R. Hunter, B. Bicknell & T. Burgess (Eds.), *Crossing divides,* (Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia), 2, 491-498 Wellington: MERGA

Steinle, V. & Stacey, K. (2012) Teachers' Views of using an on-line, formative assessment system for Mathematics. *Pre-proceedings. 12th International Congress on Mathematical Education Topic Study Group*, 33, 8 July – 15 July, 2012, (pp. 6721 - 6730). COEX, Seoul, Korea.

Steinle, V., Stacey, K. & Chambers, D. (2006). *Teaching and Learning about Decimals*. (Version 3.1). Retrieved 3 September 2015 from https://extranet.education.unimelb.edu.au/SME/TNMY/

Stewart, R., Wright, B., & Gould, P. (1998). Kindergarten students' progress in the Count Me in Too project. In C. Kanes, M. Goos & E. Warren (Eds), *Teaching mathematics in new times* (Proceedings of the 21st annual conference of the Mathematics Education Research Association of Australasia), 2, 556-563. Brisbane: MERGA

Wiliam, D. (2007). Keeping learning on track: classroom assessment and the regulation of learning. In F. K. Lester Jr (Ed.), *Second handbook of mathematics teaching and learning*. (pp. 1053 – 1098) Greenwich, CT: Information Age

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation* 37, 3–14.