Identifying Stages in a Learning Hierarchy for Use in Formative Assessment – the Example of Line Graphs.

Kaye Stacey
University of Melbourne
<k.stacey@unimelb.edu.au>

Beth Price
University of Melbourne
<b.price@unimelb.edu.au>

Vicki Steinle
University of Melbourne
<v.steinle@unimelb.edu.au>

This paper discusses issues arising in the design of questions to use in an on-line computer-based formative assessment system, focussing on how best to identify the stages of a learning hierarchy for reporting to teachers. Data from several hundred students is used to illustrate how design decisions have been made for a test on interpreting line graphs.

Designing Formative Assessment

'Smart tests' (Specific Mathematics Assessments that Reveal Thinking) are now being used by more than 500 teachers through the website www.smartvic.com. As of January 2012, there are tests on 55 topics, nearly all with paired pre-tests and post-tests. The tests are short, completed on-line, and results are immediately available to teachers. The intention is that teachers will use these tests just before teaching a topic to better understand the needs of their own students. A simple example of such use was when a teacher was preparing to teach a geometry unit to his class. In advance, his students took the understanding angle smart test which reported to him that four students had the well-known misconception that angle size is related to the length of the arms. He gave these four students a short tutorial before teaching the new topic, and was pleased that they subsequently managed the geometry unit very well. Further information on smart tests is given by Stacey, Price, Steinle, Chick and Gvozdenko (2009).

The process of creating the smart test system has raised a plethora of design issues, which have been explored through trialling with students and teachers. Issues include the nature of items suitable for on-line use; the content, scope and length of tests; the most useful content to test; the provision of information to teachers about the tests and subsequent teaching; and the method of delivery of tests to students and of student results to teachers. Many teachers have initially requested that the information they receive about their students from the smart tests system should directly link to the local curriculum standards, so that they can use this for summative (end of term) reporting to the school and parents. This implies that if teachers are to see the benefits of using formative assessment with a focus on improving day-to-day teaching, the whole system and its reporting components in particular, must be well-designed to meet their immediate teaching needs.

The aim of the paper is to describe, and illustrate with an example, the processes involved in the creation and validation of stages within a learning hierarchy which are used by the smart test system to provide teachers with information about their students.

Descriptions of Learning that are Useful for Teachers

There are many possibilities for reporting students' results to teachers. The percentage of questions correct gives teachers a ranking of students and a general idea of how well a topic has been mastered, but gives little information on what aspects need further teaching. At the other extreme, detailed reporting of every response for each individual student has the potential to provide excellent information for subsequent teaching, but there is a mass of data which takes time to interpret and does not help teachers to identify the important underlying ideas. In their own teacher-marked classroom tests, teachers have access to such

In J. Dindyal, L. P. Cheng & S. F. Ng (Eds.), Mathematics education: Expanding horizons (Proceedings of the 35th annual conference of the Mathematics Education Research Group of Australasia). Singapore: MERGA.

full information, but it has always been hard to use this other than to identify problematic items which require repeat instruction.

A different approach, which we have selected for the smart tests, is to describe learning in terms of stages along a learning hierarchy. This approach can be used with qualitative and quantitative data. The PISA study (OECD, 2010) provides an example of stages of learning derived from quantitative analysis. Using item response theory (for PISA, Rasch modelling), a scale of difficulty of the items and proficiency of students is constructed. Levels of achievement are then described by encapsulating the characteristics of the items in different regions of the scale. In PISA, proficiency is described in 6 levels and some of the most revealing country comparisons are made in terms of these levels. The level descriptions for a broad ranging survey such as PISA provide broad goals for teachers to work towards, rather than specific information on any topic. For example, OECD (2010, p.130), "At Level 6 students can conceptualise, generalise and utilise information based on their investigations and modelling of complex problem situations."

Item response theory is a major example of approaches intended to measure learning. However, many researchers interested in mathematical learning in focussed areas of the curriculum, aim to 'map' learning rather than 'measure' learning (Stacey & Steinle, 2006). This approach aims to show how ideas in a topic interact, building on each other and also interfering with each other, and to reveal the conceptions and misconceptions that underlie performance. Data to support learning hierarchies have been assembled from a wide range of sources: from written questions designed to reveal student thinking in both large-scale and small-scale studies, from task-based interviews as well as from teaching experiments. The resulting hierarchies have proved their usefulness for teaching by underpinning some significant professional development programs for teachers such as Cognitively Guided Instruction (Franke, Carpenter, Levi, & Fennema, 2001). The strategy is that teachers should know the stages of learning; know where their students are situated; and then select learning activities that will move a student forwards.

The smart tests system has adopted this general approach, describing learning in terms of a hierarchy of topic-specific stages (rather than global stages such as in PISA), reporting student performance according to the stages and then giving suggestions for how teachers can move students from one stage to the next. The descriptions of stages need to be easy for teachers to understand and easily related to teaching actions without the need for an associated professional development program. For each smart test, the report to teachers about the learning hierarchy includes additional information about specific common errors or misconceptions which are automatically diagnosed by programming based on matching patterns in student responses.

Example –Interpreting Line Graphs

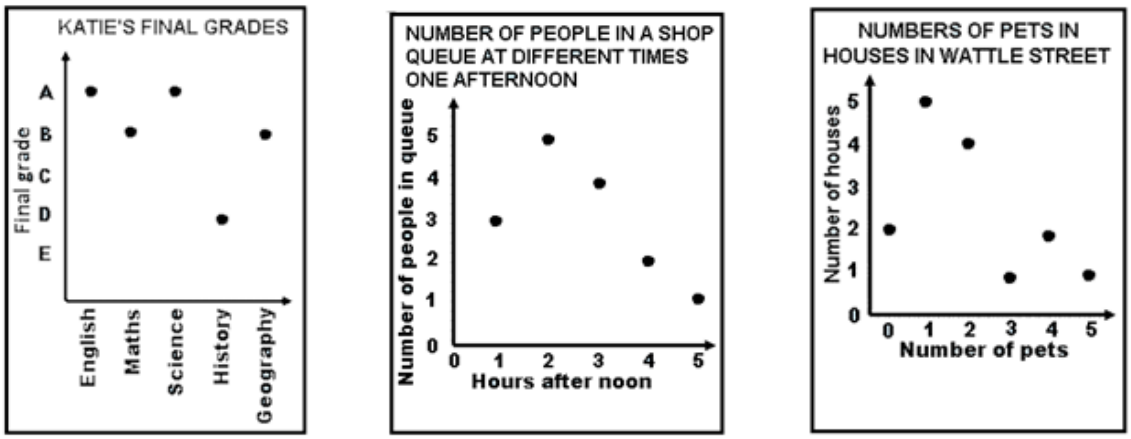
The purpose of this paper is to describe the creation of a learning hierarchy for the *interpreting line graphs* smart test. We outline the process that we have used and discuss some of the issues which have arisen. In some smart tests, the initially proposed learning hierarchy is supported by all information and data. *Interpreting line graphs* is a test for which it has been more difficult to design stages which are supported by the data. In fact, none of the hierarchies explored in this paper are of sufficient quality to be provided to teachers. So, up to the time of preparation of this paper, the reporting of students' results in this test is based only on percentage correct. The creation of this test has also underlined the often observed fact, that apparently simple topics in mathematics have interesting subtleties and hidden depths. As will be shown in the following sections, the iterative process to

identify a satisfactory learning hierarchy involves sharpening the items, reviewing the nature of the stages, and in this case, rethinking what mathematical content is fundamental.

Interpreting line graphs is one of a set of tests on the graphical presentation of data. Currently this set also includes interpreting pictographs, interpreting bar graphs, graphs: choosing the best graph type, map reference and coordinates, and plotting coordinates. There are other related tests involving, for example, scale reading. Interpretation at the higher levels requires a sophisticated degree of statistical literacy, for example, distinguishing the significance of data points and other points on a line graph.

Test creation began with the identification of the content to be tested with an indication of the likely stages of learning. Curcio's 1987 study of graph comprehension in Year 4 and Year 8 students highlighted three stages: first "reading the data" (the capacity to read literally the direct factual information on the graph); second "reading between [or within] the data" (attend to two or more data points on the graph, often for comparison purposes); and third "reading beyond the data" (extend, predict, and infer from the data). More recent work of Shaughnessy, Garfield and Greer (1996) and Shaughnessy (2007) suggests a fourth category termed "reading behind the data" which pays particular attention to the context from which the data arise. In Shaughnessy (2007) the four categories were given more detail, and expanded into eight, with the higher ones associated with deeper interpretation and appreciation of context and variation.

The *interpreting line graphs* test contains 15 items set in four contexts. Some items are multiple choice, some are short constructed response, and all are presented and marked by computer. Figure 1 contains 3 sample items. These items ask students to decide whether it makes sense to join the points in the graph, while in other items (with similar graphs) students are asked to read information from points and interpret it and also to indicate whether points intermediate between two data points have a meaning.



a) Grades b) Shop c) Pets

Figure 1. Items about joining points from the *interpreting line graphs* test. In each case, students choose Yes/No from a drop-down box after the statement "It makes sense to join these points"

Data

Data presented in this paper is from on-line users of the smart tests in 2011, so they are the students of volunteer teachers from many schools. The 216 students are from Years 5 to 10, with 59% from Year 8. Data collected before 2011 was used to identify features of the task that are significant for students, improve the items, and trial the early version of the reporting system. Further refinement of the test and reporting is now taking place.

Designing and Confirming the Learning Hierarchy

Version 1 of the Learning Hierarchy

A learning hierarchy can be created empirically from data and/or by postulating a complexity order based on logical analysis, and/or teaching experience and/or using prior research (e.g. in this case, Curcio, 1987; Shaughnessy, Garfield and Greer, 1996 and Shaughnessy, 2007). Stages in a learning hierarchy are confirmed by data if several conditions are met. Ideally, items with similar mathematical characteristics will have similar success rates, and will be completed successfully by the same students. If 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 unlikely to successfully complete items designed to test higher stages. Moreover they should have a low 'relative risk' of completing the item successfully, where relative risk is the ratio of probability of success of students at a given stage to the probability of success for all students.

The first attempt to identify a learning hierarchy is shown (with brief examples) in Table 1. It is based on the four categories of Shaughnessy and colleagues, applied specifically to line graphs and expressed in concrete terms. Several items are created for each stage (referred to as, for example, *Stage 1 items*); a hurdle score on a group of items is set (for example, 4 out of 5 correct is the requirement for success on the Stage 1 items); and then students are classified according to their performance on the groups of items (for example, being successful on Stage 1 items followed by not being successful on Stage 2 items leads to a student being classified as Stage 1).

Table 1
Version 1 of the Learning Hierarchy for Interpreting Line Graphs

Stage	Description: Students can	Brief example	Number of items
1	Read information from points on labelled grid lines	Read Katie's grade in English (Fig. 1a)	5
2	Read information from points between labelled grid lines	Read the weight of a mouse from axis marked in 5 gram intervals	3
3	Interpret the information in a line graph	Identify when a mouse had babies from a graph of daily weight.	2
4	Decide whether it is appropriate to join data points	See 3 items in Figure 1	5

The data from the full set of 15 items was analysed using the stages in Table 1. The first column of Table 2 shows that, of the 216 students, 49 did not reach Stage 1 (so are classified as Stage 0), 28 were classified as Stage 1, 99 as Stage 2, 39 as Stage 3 and only 1 was classified as Stage 4. The entries in the main cells of Table 2 are the average probabilities of success (and in brackets, relative risk) for the items contributing to each stage, for the groups of students who have been classified into each stage.

Table 2
Average Probability of Success (relative risk) on Items for Students Allocated to Each Stage in Version 1 of the Learning Hierarchy

Classification of students	Items in each stage			
	Stage 1 (5 items)	Stage 2 (3 items)	Stage 3 (2 items)	Stage 4 (5 items)
Stage 0 (n = 49)	0.32 (0.40)	0.27 (0.42)	0.18 (0.48)	0.35 (0.88)
Stage 1 (n = 28)	0.91 (1.14)	0.27 (0.43)	0.36 (0.92)	0.45 (1.12)
Stage 2 (n = 99)	0.95 (1.18)	0.84 (1.33)	0.25 (0.64)	0.40 (1.01)
Stage 3 (n = 39)	0.97 (1.24)	0.82 (1.30)	1.00 (2.59)	0.41 (1.02)
Stage 4 (n = 1)	1.00 (1.24)	0.67 (1.06)	1.00 (2.59)	0.80 (2.00)
All students (n = 216)	0.80	0.63	0.39	0.40

Table 2 shows that the data supports Stages 1, 2, and 3 moderately well. For example, inspection of the relative risks in the columns related to the items in Stages 1, 2 and 3 reveals that the cells above the main diagonal are less than 1 and the cells under the main diagonal are greater than 1. Also, in the columns related to Stages 1, 2 and 3 in the upper triangle, the average probabilities of success are low (no more than 0.36). However, we would expect increasing trends down the columns and this is not the case; for example, compared to the Stage 1 students, Stage 2 students have *less success* on the Stage 3 and 4 items. Stage 4 (knowledge of when it is appropriate to join points and the meaning of interpolated points) is especially problematic. The final column in Table 2 shows that students at Stages 1, 2 and 3 all have similar average probabilities of success on these 5 items. In summary, this data does not support these four stages as a learning hierarchy.

Version 2 of the Learning Hierarchy

In an attempt to resolve the issues with the (linear) learning hierarchy in version 1, a branching hierarchy was proposed. The original Stages 1, 2 and 3 in version 1 (reading and interpreting points) remain, but knowledge of when it is appropriate to join points in a graph (the original Stage 4) would not follow on from Stage 3, but rather branch from Stage 1. Table 3 indicates that 5 items were used to determine Stage 1 (the original Stage 1 items) and another 5 items (the original Stage 4 items in version 1) were now used to classify students into Stages 2, 3 and 4 in this new branch of the learning hierarchy. This decision was supported by the use of Statistical Implicative Analysis performed by the CHIC software (Gras 2010) which confirmed that these 5 items about joining points differed from the other items in more than just degree of difficulty.

Table 3 provides the success rates for each of the 10 items for the full cohort of 216 students; the decreasing trend providing initial support for this new branch in the learning hierarchy. The 3 items defining Stage 2 and 3 are in Figure 1. The differences between the success rates of the shop item (71%) where it is appropriate to join the points and the other two items where it is not appropriate (45% and 59%) contributed to the decision to create separate stages.

The two Stage 4 items in version 2 addressed this issue more deeply, by asking for the meanings of interpolated points. Sometimes such points are meaningless but for graphs with continuous data on both axes it is usually possible to interpolate some information. For example, one Stage 4 item involved an altered version of the graph in Figure 1b where the

points were joined and an interpolated point (at 3:30pm) was marked. In this *joined-points shop item*, students were asked to select a suitable meaning for this marked point from this list:

A) that 35 people were in the shop at 3 pm
B) that 30 people were in the shop at 3:30 pm
C) that you could expect about 30 people in the shop at 3:30 pm (Correct)
D) nothing. This point does not have a meaning.

The other Stage 4 item (the *joined-points pet item*) was derived in a similar way from Figure 1c, with an interpolated point (2.5, 2.5) marked. Again, students were asked to select a suitable meaning for this marked point from a list:

A) that 2.5 houses had an average of 2.5 pets each
B) that house 2A had 2.5 pets
C) nothing. This point does not have a meaning. (Correct)

Table 3
Version 2 of the Learning Hierarchy for Joining Points within Interpreting Line Graphs

Stage	Description: Students can ...	Number of items used	Success rates for each item (n=216)
1	read and interpret data points	5	83%, 75%, 83%, 84%, 76%
2	decide to join points	1	71%
3	reject inappropriate joining of points	2	45%, 59%,
4	interpret interpolated points	2	13%, 12%

Table 4 provides data for the version 2 learning hierarchy in the same way that Table 2 provided data for version 1.

Table 4
Average Probability of Success (relative risk) on Items for Students Allocated to Each Stage in Version 2 of the Learning Hierarchy for Joining Points

Classification of students	Items in each stage			
	Stage 1 (5 items)	Stage 2 (1 item)	Stage 3 (2 items)	Stage 4 (2 items)
Stage 0 (n = 49)	0.32 (0.40)	0.57 (0.81)	0.42 (0.80)	0.17 (1.39)
Stage 1 (n = 42)	0.95 (1.18)	0.00#(0.00)	0.36 (0.68)	0.13 (1.05)
Stage 2 (n = 77)	0.94 (1.17)	1.00 (1.41)	0.38 (0.73)	0.10 (0.83)
Stage 3 (n = 47)	0.96 (1.19)	1.00 (1.41)	1.00 (1.91)	0.09 (0.68)
Stage 4 (n = 1)	1.00 (1.24)	1.00 (1.41)	1.00 (1.91)	1.00 (8.00)
All students (n = 216)	0.80 (1.00)	0.71 (1.00)	0.52 (1.00)	0.13 (1.00)

This is an artefact of having only one Stage 2 item.

While the relative risks in the lower triangle in Table 4 are larger than 1 (as is desirable), not all of those in the upper triangle are less than 1. The relative risk values of 1.39 and 1.05 indicate that Stage 0 and Stage 1 students are more likely to answer the Stage 4 items correctly than average, which does not fit a learning hierarchy. Students are classified at Stage 0 if they did not show adequate skills reading single points, so it is rather surprising that they were more successful than other students on the Stage 4 items. It seems that the

students with higher scores were especially reluctant to say that a point on an interpolated line has no meaning. For example, in the joined-points pet item they were much more likely to choose response A (2.5 houses have an average of 2.5 pets) than response C (This point does not have a meaning). They were keen to show the ability to read the coordinates points, rather than the ability to 'read behind the data'.

Rethinking the Mathematical Content and the Items

The current test is not producing results that are easy to analyse and to report meaningfully to teachers. We have therefore undertaken a major revision, with several steps. First, some items have been subjected to minor editing, usually to improve clarity. Wording and graphs can often be clarified and extraneous cognitive load can be removed. For example, the unnecessary word 'queue' has been removed in the new version of the shop item (shown in Figure 1b). Even though minor edits such as these make little difference to overall success rates, they may help to sharpen the data from individuals, reducing 'careless errors' which produce noise in the data. Examining the wrong answers that students have given, also gives clues for minor improvement. For example, in the graph showing daily measurements of the mass of a mouse, many students were out by exactly one day and hence we have clarified the time of day when the mouse was weighed.

The second aspect to the revision is to rethink what should be tested by this test. Mathematics is a highly connected subject, so it is hard to draw strong boundaries around any topic. On reconsidering this test, it seems that one possibility for the difficulty in producing a learning hierarchy is that the test has had too many mathematical facets. The test had also incorporated some scale reading features, but this is not necessary as there is another smart test that looks specifically at reading scales. To avoid these scale features, all points on the new graphs have been placed at the intersection of labelled grid lines.

Reviewing the Fundamental Ideas

A third aspect is to rethink what are the fundamental mathematical ideas for this test. In the first version, we focussed on recognising when to join or not join data points and we asked students to give the meaning (or non-meaning) of data points. However, these items gave unreliable data. We now think that there are two competing considerations for students seeking to make this decision, and we had not taken this well enough into account. The first is 'Does joining the points make it easier to read the information on the graph by guiding the eye?' and the second (more mathematically sophisticated, and the only intention of the initial items) is 'Does it make sense to use this line for interpolation?' Having considered these issues, it now seems to the authors that it is almost always easier to read a set of data points when they are joined. In effect, we now agree with most students that all of the scattergraphs in Figure 1 are easier to read when the points are joined. So the fundamental issue of joining points is not whether points should be joined, but what additional information (if any) can be gained by looking at points on the interpolating line. This requires 'reading behind the data' taking the real context into account.

In the redesigned test, questions are again asked about data points and interpolated points, but focussing on the status of information. For example, in the new joined-points shop item, students will be asked to select whether the statement 'At 3.30 pm there were about 30 people in the shop' is true / likely / unlikely / not true / the graph gives no information on this. Overall the new version of the test should produce data that can more readily be reported to teachers on one learning hierarchy, which is better focussed on the real issues of reading and interpreting line graphs.

Conclusion

The creation of a learning hierarchy is an iterative process, cycling between consideration of mathematical goals and potential items, the collection of data, and the analysis both of overall success rates and the examination of student errors in order to gain insight into both student thinking and ways of improving the items.

Reporting results in a learning hierarchy is attractive if the hierarchy description helps teachers focus on the achievements of their students and also what they need to master next. However, as the example above shows, it is not always easy to design tests which fit the empirical and theoretical criteria well. We have not yet trialled our new test, but hope that having built it on the basis of data collected to date, it will have good statistical properties.

On-line testing and computer diagnosis form only a part of assessment of mathematics; although there are new opportunities for item types the possibilities for constructed responses are still limited. On-line testing does, however, through systems such as the smart tests, offer new opportunities to mobilise the results of mathematics education research, making it available to teachers at the point of teaching.

Acknowledgements

We thank Eugene Gvozdenko, Reinhard Oldenburg, Helen Chick, teachers and students, the Australian Research Council and DEECD of Victoria.

References

Curcio, F. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education, 18*, 382-393.

Franke, M., Carpenter, T., Levi, L., & Fennema, E. (2001). Capturing Teachers' Generative Change: A Follow-up Study of Professional Development in Mathematics. *American Educational Research Journal, 38*(3), 653 – 689.

Gras, R. (2010): CHIC 5.0 (Software). http://www.amarkos.gr/en/research/chic/

OECD (2010). *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Vol I)*. Paris: Organisation for Economic Cooperation and Development. http://dx.doi.org/10.1787/9789264091450-en

Shaughnessy, J. M., Garfield, J., & Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics in education* (pp. 205-237). Dordrecht, The Netherlands: Kluwer.

Shaughnessy, J. M. (2007) Research on statistical learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp.957-1009). Charlotte, NC: Information Age Publishing.

Stacey, K. & Steinle, V. (2006). A case of the inapplicability of the Rasch Model: Mapping conceptual learning. *Mathematics Education Research Journal, 18*(2), 77 – 92.

Stacey, K., Price, B., Steinle, V., Chick, H., Gvozdenko, E. (2009). SMART Assessment for Learning. *Proceedings of International Society for Design and Development in Education Conference*. http://www.isdde.org/isdde/cairns/pdf/papers/isdde09_stacey.pdf