KEYNOTE

Rethinking measurement for accountable assessment

Professor Mark Wilson
University of California, Berkeley & University of Melbourne

https://doi.org/10.37517/978-1-74286-638-3_13

Mark Wilson is a professor of education at the University of California, Berkeley, and at the University of Melbourne. He teaches courses on measurement in the social sciences, especially as applied to assessment in education. He was elected President of the Psychometric Society, and, more recently, President of the National Council for Measurement in Education (NCME). His research and development interests focus on the development and application of sound approaches for measurement in education and the social sciences, the development of statistical models suitable for measurement contexts, the creation of instruments to measure new constructs, and scholarship on the philosophy of measurement.

Abstract

The underlying model for most formal educational measurement (e.g. standardised tests) is based on a very simple model: the student takes a test (possibly alongside other students). The complications of there being an instructional plan, actual instruction, interpretation of the outcome, and formulation of next steps, are all bypassed in considering how to model the process of measurement. There are some standard exceptions, of course: a pre-test/post-test context will involve two measurements, and attention to gain score, or similar. However, if we wish to design measurement to hold to Lehrer's (2021) definition of 'accountable assessment' – as 'actionable information for improving classroom instruction' – then this narrow conceptualisation must be extended. In this presentation, I will posit a simple model that reflects the simple one-test context described above, and then elaborate on it by adding in a) a framework for design of the assessments that is keyed to educational interpretation, b) further rounds of data collection that can indicate changes in a student's underlying ability, and c) provision for varied assessment modes that will allow for i) classroom-independent tasks that operate at the summative and meso levels, and ii) classroom-dependent tasks that operate at the micro level. The former are designed to provide a basis for triangulating student responses across different contexts, and the latter are designed to closely track the variation of student performance over time in a classroom instructional context. This framing will be exemplified in a in a K–5 elementary school that is seeking to improve the quality of instruction and students' understandings of measure and arithmetic. The different levels of data collection will be instantiated by two different pieces of software, which operate at the micro level and the meso/summative levels respectively.

Introduction

The underlying educational context for educational measurement has generally been one where a student is seen as progressing through an instructional plan delivered by their teacher(s); the student (along with their peers) is then subjected to a test designed to assess the expected range of outcomes, and the teacher(s) then plan for the next instructional step based on that assessment.

Of course, there may be further rounds of this – retesting, etc. Although this rather general formulation is well-known to most involved in measurement, the actual paradigmatic context on which educational measurement is predicated is much simpler – the student takes a test (possibly alongside other students). The complications of there being an instructional plan, actual instruction, interpretation of the outcome, and formulation of next steps, are all bypassed in considering how to model the process of measurement. There are some standard exceptions, of course – a pre-test/post-test context will involve two measurements, and attention to gain score, or similar.

We see these types of testing as being representative of the macro level of assessment. The *macro level*, commonly also called 'summative testing', is the level of most standardised tests. The topics are relatively coarse composite constructs, such as science, language arts, geometry, and the tests are largely used for administrative decisions by parents and administrators/policymakers. They are used over relatively longer education time-periods (years, semesters, program length, etc.) for relatively large-scale decision-making, such as passing a course, grade advancement and course-placement.

However, if we wish to design measurement to hold to Lehrer's (2021) definition of *accountable assessment* as 'actionable information for improving classroom instruction', then this narrow conceptualisation must be extended. In this presentation, I will posit a model that starts with the simple one-test context described above, and then elaborate it by adding in 1) a framework for design of the assessments that is keyed to educational interpretation, 2) further rounds of data collection that can indicate changes in a student's underlying ability, and 3) provision for varied assessment modes that will allow for a) classroom-independent tasks that operate at the macro (summative) and meso levels, and b) classroom-dependent tasks that operate at the micro level. The *meso level* assessments are designed to provide a basis for triangulating student responses across different contexts, and the *micro level* assessments are designed to track closely the variation of student performance over time in a classroom instructional context.

This framing is exemplified in work from a K–5 elementary school that is seeking to improve the quality of instruction and students' understandings of measure and arithmetic. The different levels of data collection will be instantiated by two different pieces of software, which operate at the micro level and the meso/macro levels respectively.

The National Science Foundation Collaborative Research Project *Modeling assessment to enhance teaching and learning* (MAETL) is a collaboration among Richard Lehrer, Leona Schauble and Corey Brady from Vanderbilt University, and Mark Wilson and Perman Gochyyev from the University of California, Berkeley. The purpose is to create and test-out a novel assessment system designed to address two coordinated purposes:

1. to provide ongoing, instructionally productive evidence to teachers about student learning in the context of learning progressions
2. to link dense information from student work products and formative assessments (meso and micro assessments) in new models that generate robust estimates of the growth of student learning.

The specific topics of instruction are Measurement of Length, Area, Volume, Angle, and Measured Quantities (as entrée to Rational Numbers – Fractions as quantities, fractions as operators). Consider the first construct, Theory of Measurement – Length (ToML) as an example; this construct describes how children come to constitute a theory of measure to compare magnitudes (extents) of lengths. and is represented using a 'construct map', based on the Bear Assessment System (BAS; Wilson, 2005) and is illustrated in Figure 1. The levels of a construct map are designed to encapsulate important qualitative steps towards the highest level.

For example, at the second level, ToML 2, students focus on the nature of a unit. They must learn (in practice) that:

1. units enable indirect measurements via accumulation and count (instead of directly comparing), and
2. units allow for both additive (how much longer?) and multiplicative comparisons (how many times longer?).

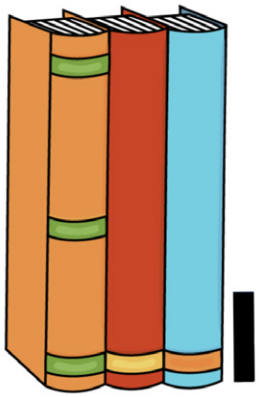
Figure 1 The construct map for TOML

6. Generalising relationships (e.g. Measure of A in B is the reciprocal of measure of B in A)
5. Partitioning and symbolising involving 3-splits and composition of 2- and 3-splits
4. Partitioning, iterating, symbolising partitioned units – 2-splits
3. Iterating units and symbolising distance travelled
2. Explaining properties of units and their role in accumulation
1. Directly comparing

Students must develop understandings of the properties that enable these uses. Hence, students at this level need to be able to explain the roles of identical units and tiling. Then, as they move up to ToML 3, they must show that they can use these units to measure something, such as is exemplified in the item shown in Figure 2. The items are developed and delivered using the BEAR Assessment System Software (BASS; Wilson et al., 2019).

Figure 2 An item at level 3: Height of a Book

Carlos started measuring the height of the blue book, but he does not have enough units. Carlos says he cannot finish measuring the height.

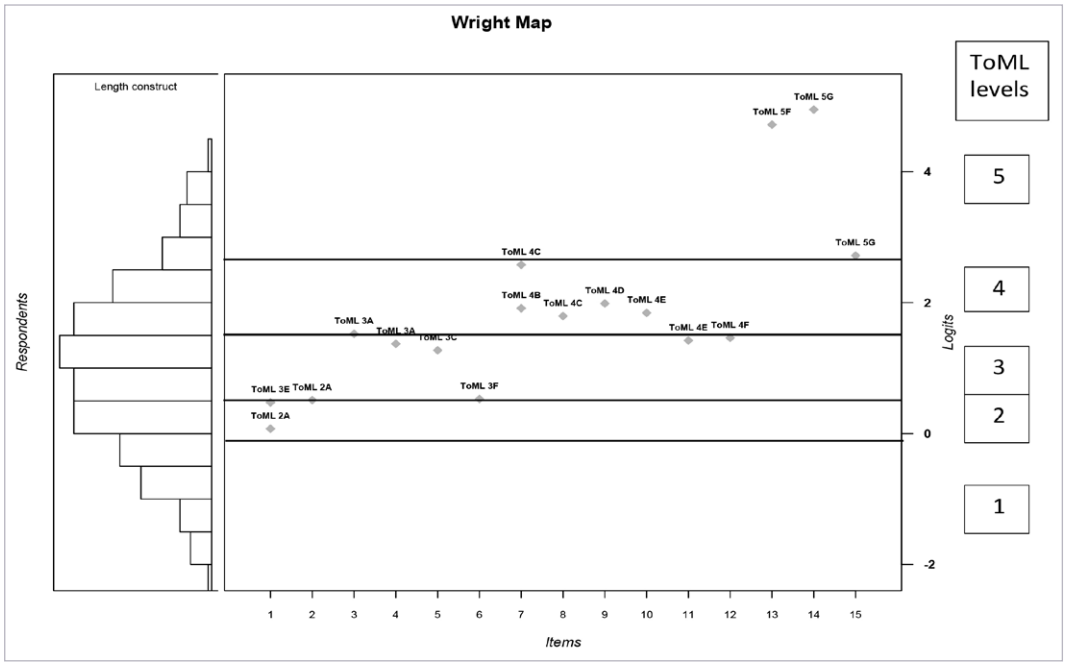


Do you agree with Carlos?

- ○ Yes
- ○ No

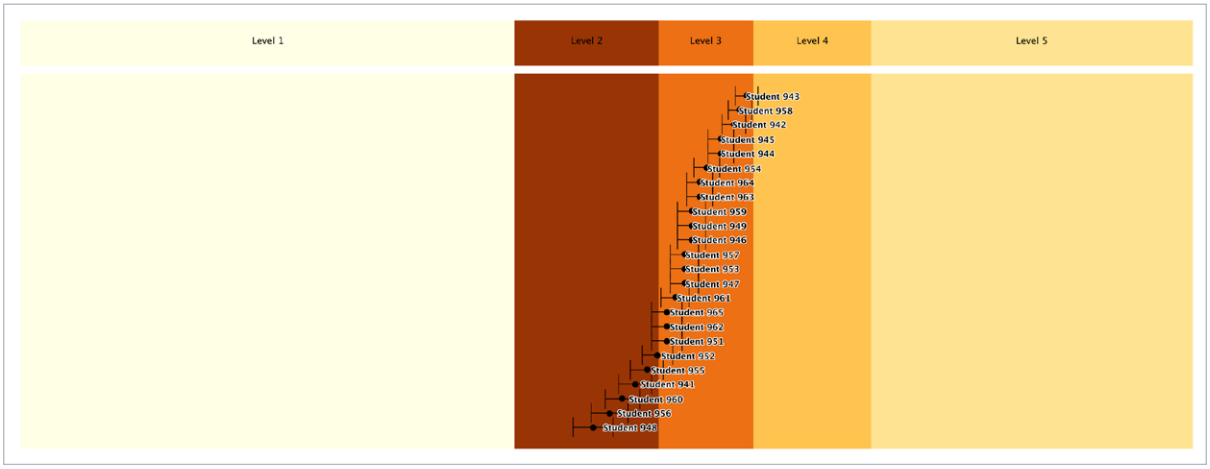
Items such as these are designed to be used at the meso level. This is the level of testing for teaching and tends to deal with the contents of broader standards. Specifically, it relates to what one might call 'teachable constructs', such as, buoyancy, variability or measurement. They can be used by teachers to put together tests, and the results would be used by teachers, and, when old enough, students themselves. Typically, they might be used at instructional-scale time-periods and up (i.e. days, weeks, teaching units, etc.) for making instructional decisions such as planning for a day/week's topics, what topics to revisit, which students need extensive help. Results for the ToML items such as these can be displayed by BASS as Wright Maps, and the locations of the item thresholds can be graphically summarised. A Wright Map is a graph that simultaneously shows estimates for both the students and items on the same (logit) scale, an example of which is shown in Figure 3. On the left side of the Wright Map, the distribution of student abilities is displayed as a histogram (on its side), where ability entails knowledge of the skills and practices for ToML. The person's abilities have a roughly symmetric distribution. On the right side are shown the thresholds for 15 items in ToML (where two of the items are polytomous – that is, the student responses can indicate more than one level of the ToML construct map). Each item is represented as 'ToML Lk', where L represents the ToML Level of that item, and k represents a finer-grade coding not discussed in this brief paper. Looking beyond the results for a single item, we note the consistency of the locations of these thresholds across items. We used a standard-setting procedure called 'construct mapping' (Draney & Wilson, 2011) to develop cut-scores between the levels. Following that process, we found that the thresholds fall quite consistently into the ordered levels from ToML 1 to ToML 5, with a few exceptions, at the borderlines between the levels.

Figure 3 Wright Map for Theory of Measurement – Length



Having established these cut points between ToMl Levels, we also can use purpose-built BASS reports, such as the Group Report shown in Figure 4, which lay out the distribution for a whole class across the ToML Levels 1 to 3. In these reports, the estimated student location is noted as a black dot, and a 66.7 per cent (i.e. 1-sigma) confidence region is shown around that location. Here we can see that this class extends from Level 2 to Level 3, just bordering on Level 4.

Figure 4 Group report for a (fictional) class



The construct map can be used to tie in the meso level assessments with the micro level in the classroom. At the micro level, relatively informal and fine-grained assessments are used for within-instruction observations. These typically relate to relatively fine-grain sized knowledge, 'in-pieces' such as the definition of density, initial experiences of variation, what it means to measure a length, etc. They are associated with the opportunities teachers have for telling observations as they walk around their classrooms and are focused on brief education time-periods (i.e. a sequence of concepts within a classroom unit, etc.), and are intended to inform smaller-scale instructional decisions, such as what tactic to use next in discussing a certain idea with students. If one were to be developing computerised teaching software, then this is essentially the same level as the software would be operating at.

The levels of the construct map are the key to bringing these two levels of information together. The project has developed Teacher Observation Tools (TOTs), which is a mobile data-gathering iPad application designed for teacher use while teaching in their classrooms. A sample screenshot is shown in Figure 5. To use this software, a teacher must learn the construct map (in this case the ToML construct map) well enough to apply it 'on the fly' in their classroom – note the selection of ToML3A for the recorded classroom event. An example of the data records for one classroom is shown in Figure 6. These data records can be input to the BASS database and a sample group report is illustrated in Figure 7.

Figure 5 A screen shot from TOTs

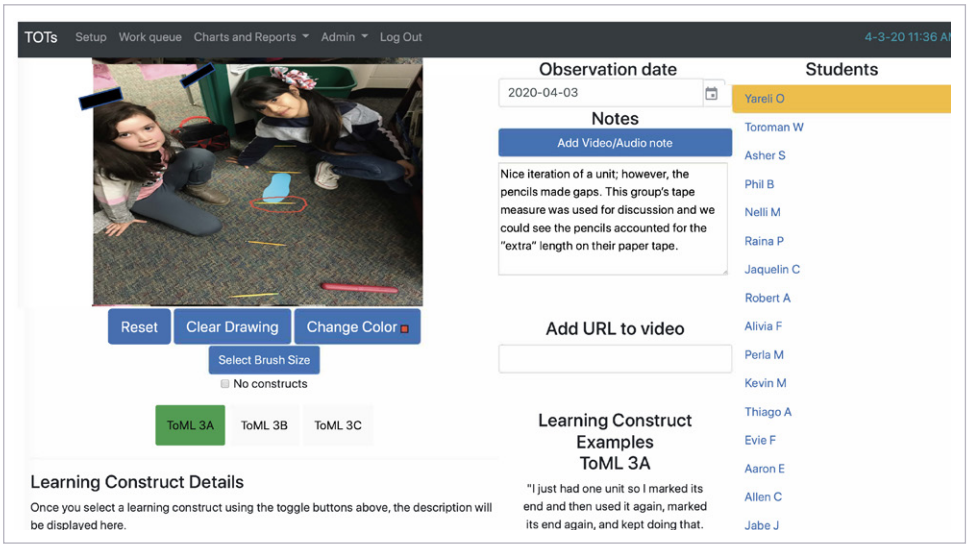


Figure 6 TOTs record for across time for one class

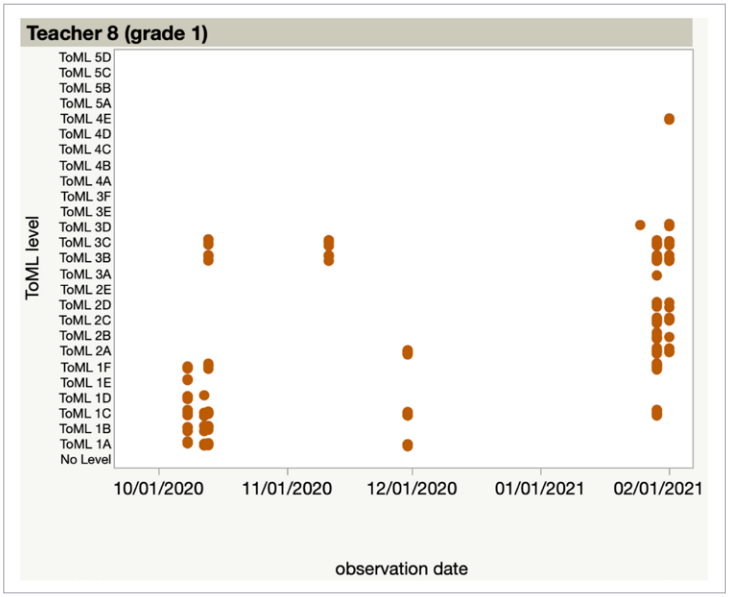
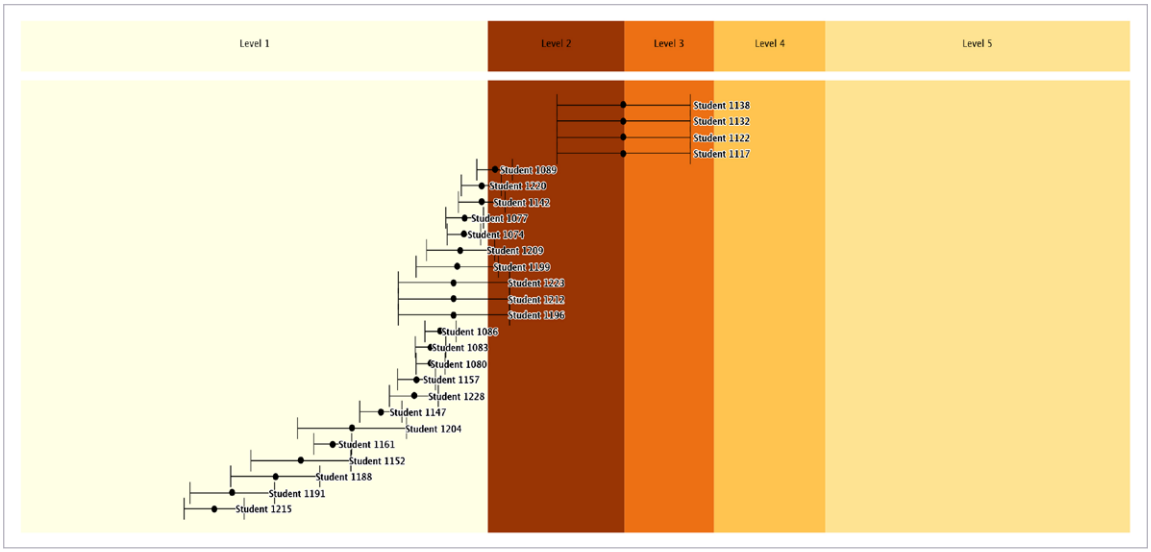


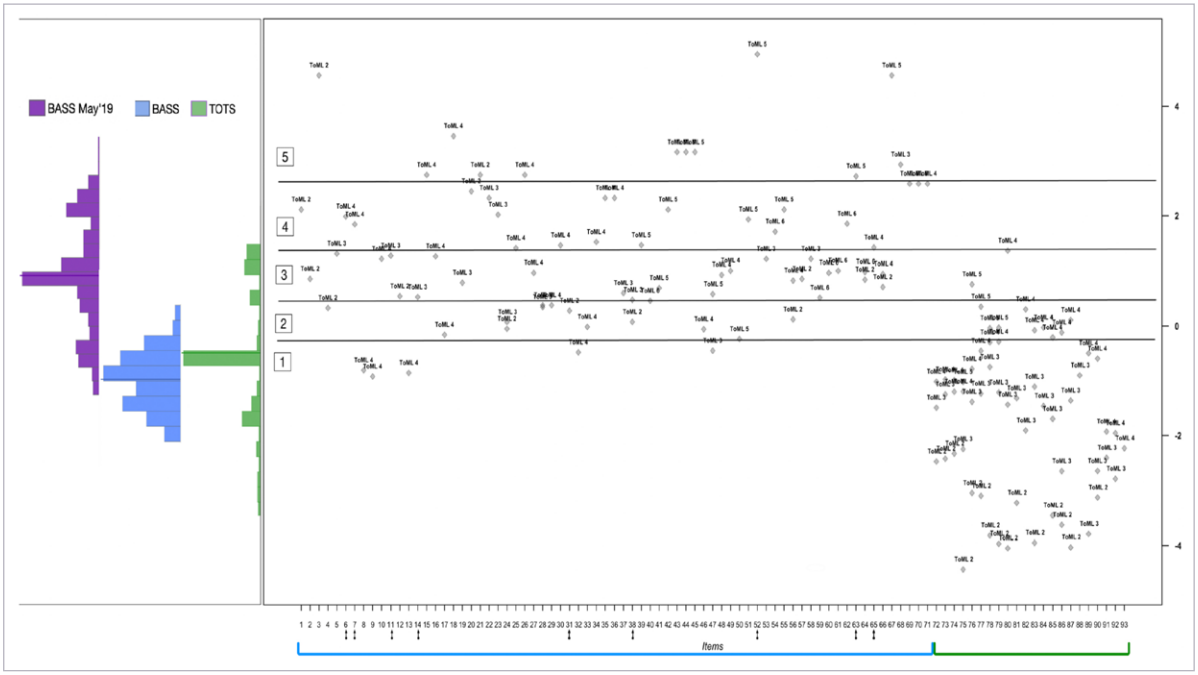
Figure 7 BASS group report for the TOTs data



The two levels are coordinated via the levels of the ToML construct map, and representative results from scaling the two together are shown in the expanded Wright Map in Figure 8, where the BASS levels illustrated in Figure 3 are used to interpret the findings. These ToML Levels were established in the 2019 post-test, for which the student estimates are shown on the left-hand side in the purple histogram. As we can see, the students, who ranged from Grade 1 to Grade 5, spread along the scale from ToML Level 1 to 5. Next to the right is a blue histogram showing the student estimates at the pretest in 2020 – as one might expect, these are much lower than for the previous year's post-test results, ranging up to only Level 2. The next histogram (green) shows the TOTs estimates for students in the first part of the 2020/21 academic year (i.e. the same students as shown in the blue histogram), and here we see an interestingly broader range than for the pretest. This increase in breadth can be attributed partly to 1) the initial instruction in the program, but also to 2) the scaffoldings to student performance provided in the classroom, and 3) the increased appreciation of teachers for the communications of their own students. One interesting extra feature is the locations

of the TOTs items on the right-hand side of this Wright Map: first, it is shown as a 'cloud' of micro-level items, as we modelled them as random-effects, and second, the locations are all lower than for the meso-level BASS items, and we see this as being attributable to, again, 1) the scaffoldings to student performance provided in the classroom, and 2) the increased appreciation of teachers for the communications of their own students.

Figure 8 Reconciling the meso and micro levels of results



In conclusion, we note that the conceptualisation of 'accountable assessment' involves the matching development of assessment at both the meso and micro levels of assessment, as exemplified in this brief paper. The BASS and TOTs software accommodate these assessment levels, and the possibility of coordinating between the two is accomplished by basing *both* on the relevant construct map (in this case ToML). Other approaches to this conceptualisation are also possible (e.g. Doignon & Falmagne, 1999), though generally they are built only at the finest (micro) grain size, which, while needed for applications such as computer-based teaching, may not serve human teachers so well.

References

Doignon, J. P. Falmagne, J. C. (1999). *Knowledge Spaces*. Springer-Verlag.

Draney, K., & Wilson, M. (2011). Understanding Rasch measurement: Selecting cut scores with a composite of item types: The Construct Mapping procedure. *Journal of Applied Measurement*, *12*(3), 298–309.

Lehrer, R. (2021, August 16–20). *Accountable assessment*. [Keynote presentation]. Australian Council for Educational Research Conference (online).

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Erlbaum.

Wilson, M, Scalise, K., and Gochyyev, P. (2019). Domain modelling for advanced learning environments: the BEAR Assessment System Software. *Educational Psychology*, *39*(10), 1199–1217. https://doi.org/10.1080/01443410.2018.1481934.