Metrological traceability in education: A practical online system for measuring and managing middle school mathematics instruction

D Torres Irribarra[1,2], R Freund[1], W Fisher[1], M Wilson[1]
[1] BEAR Center, Graduate School of Education, University of California, Berkeley, CA (USA)

Abstract. Computer-based, online assessments modelled, designed, and evaluated for adaptively administered invariant measurement are uniquely suited to defining and maintaining traceability to standardized units in education. An assessment of this kind is embedded in the Assessing Data Modeling and Statistical Reasoning (ADM) middle school mathematics curriculum. Diagnostic information about middle school students' learning of statistics and modeling is provided via computer-based formative assessments for seven constructs that comprise a learning progression for statistics and modeling from late elementary through the middle school grades. The seven constructs are: Data Display, Meta-Representational Competence, Conceptions of Statistics, Chance, Modeling Variability, Theory of Measurement, and Informal Inference. The end product is a web-delivered system built with Ruby on Rails for use by curriculum development teams working with classroom teachers in designing, developing, and delivering formative assessments. The online accessible system allows teachers to accurately diagnose students' unique comprehension and learning needs in a common language of real-time assessment, logging, analysis, feedback, and reporting.

1. Introduction

Traceability to consensus standard units is a primary means of coordinating quantitative communications in science [1]. In any given field, theory, data, and instruments focused on phenomena expressed in common terms improve the likelihood of effectively applying new results, of successfully advancing the state of the art, and of obtaining significant economic efficiencies [2]. The third edition of the International Vocabulary of Metrology [3] therefore asserts the general value of standardized terms and concepts, including psychological and social measurement for the first time.

Of course, many more questions than answers follow from the suggestion that some form of recognizably metrological concepts and methods might be achievable in psychology and the social sciences. Many of the problems center on connecting measures of additive effects in a common framework that supports data- and theory-based checks on the meaning and quality of the unit. Building on advances in theory and method [4], research conducted over the last several decades has identified, modelled, scaled, and described the substantive properties of a number of invariant psychological, social, and physical constructs [5-7]. But a variety of reasons, many involving assumptions about the nature of quantification and the culture of measurement practices in education

[2] To whom any correspondence should be addressed.

and psychology, have prevented development and tests of viable methods for effectively and efficiently maintaining a shared metrological frame of reference for the constructs of interest in these fields. Work that can be broadly construed as relevant to this purpose has largely involved computerized administration of item banks, which lends itself to implementation in networked contexts.

However, online test administration alone is an insufficient response to the challenges that must be faced. Recent experience has decisively demonstrated the dysfunctional effects caused when summative accountability purposes are imposed on educators who do not have access to the tools they need if they are to take responsibility for their students' outcomes [8]. Combining summative and formative applications is complex [9]. Effective assessment of learning readiness, and putting the results of those assessments to use in the classroom, can be incorporated in systematic routines to limited extents. The most challenging difficulties are those involving close and ongoing (formative) attention to individual differences in students, curricula, teachers, schools, which also require the monitoring of overall (summative) progress. The online environment provides a natural way to provide the tools needed to guide teachers and others in the performance of complex operations they would not otherwise attempt. Framing the cognitive context for asking the right questions is important for measurements to inform immediate instructional goals, as well as to assess longer term outcomes. Innovative technologies and advances in predictive theoretical models are now rapidly changing the paradigm for what is possible in instructionally-oriented assessment [10-11].

2. Sound measurement for quality assessment

The capacity to bring the right questions, technologies, and information together in the education context is the focus of the BEAR Assessment System [12], a measurement approach built on the idea that quality assessment demands sound measurement. This system emphasizes a cyclical process of construct (measurand) characterization, item development, outcome definition and evidence collection based on analysis of student responses (Fig. 1). This process is aimed at providing meaningful interpretations of student work relative to the cognitive and developmental goals defined by a curriculum.

The BEAR Assessment System is built on four principles:

1. A developmental perspective that specifies what is being assessed and how it will be identified, from emerging perspectives to the developing views of higher order proficiency.
2. A match between instruction and assessment that ensures what is taught is what is measured, and that teaching, learning and assessment support each other in improving student learning.
3. Appropriate feedback, feed forward, and follow-up that engage teachers and the students in the assessment process and that brings metacognition fully into play to improve outcomes.
4. Generation of quality evidence, so that teachers, parents, schools and other stakeholders can have confidence in reported outcomes, and can readily interpret and use assessment findings to improve instruction and student learning outcomes.

The Berkeley Assessment System Software (BASS; Fig. 2) is a web application implementing the BEAR Assessment System. It allows researchers and teachers to design, develop, and deliver formative assessments and to monitor and report student progress within an interpretive context. BASS is being developed to support the use of the BEAR Assessment System. Accordingly, BASS is designed to be used in close alignment with educational goals and instructional content. This allows teachers to understand the assessment results in terms of the learning progressions defined by each curriculum, and to diagnose students' comprehension levels and learning needs according to their placement along these progress variables.

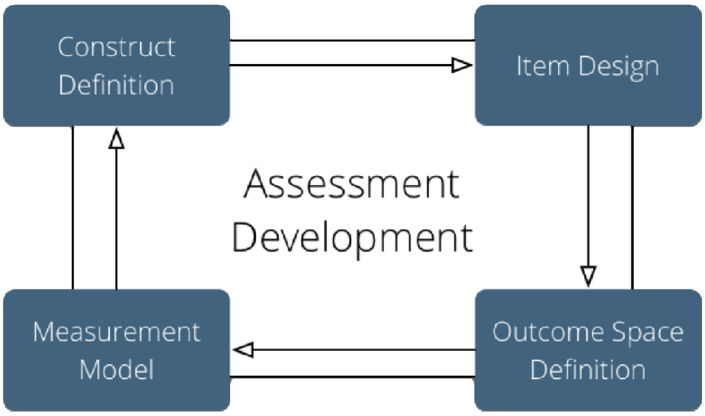


Fig. 1. The BEAR Assessment System

This system allows easy management of all kinds of assessment, including in-class activities, homework assignments, end-of-unit tests, and more. BASS builds on more than a decade of research in the design of assessment tasks and measurement techniques to support evidence-based assessment in the classroom [13-15]. In order to make the system usable to instructors in classrooms, teachers are placed at the center of the design process. The development of the teacher tools in BASS emphasizes (1) giving teachers flexibility in designing and assigning activities for students, (2) simplifying the scoring process, and (3) providing them with useful reports to share with students and parents and to use in planning instruction. The assessments include computerized modifications of existing paper-and-pencil tasks; new computer-based assessment tasks (e.g., card-sort problems, interactive graph problems) are planned for future development. An estimation engine under development will automate current manually conducted comparisons of student performances on paper-and-pencil and computer-based tasks at both the task and test level.

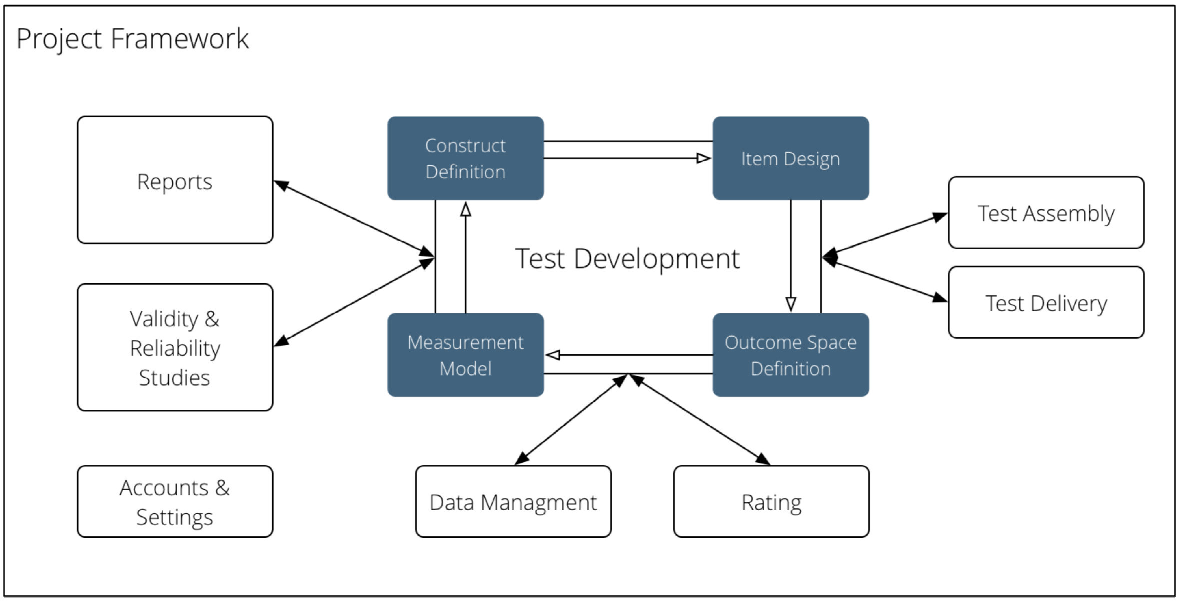
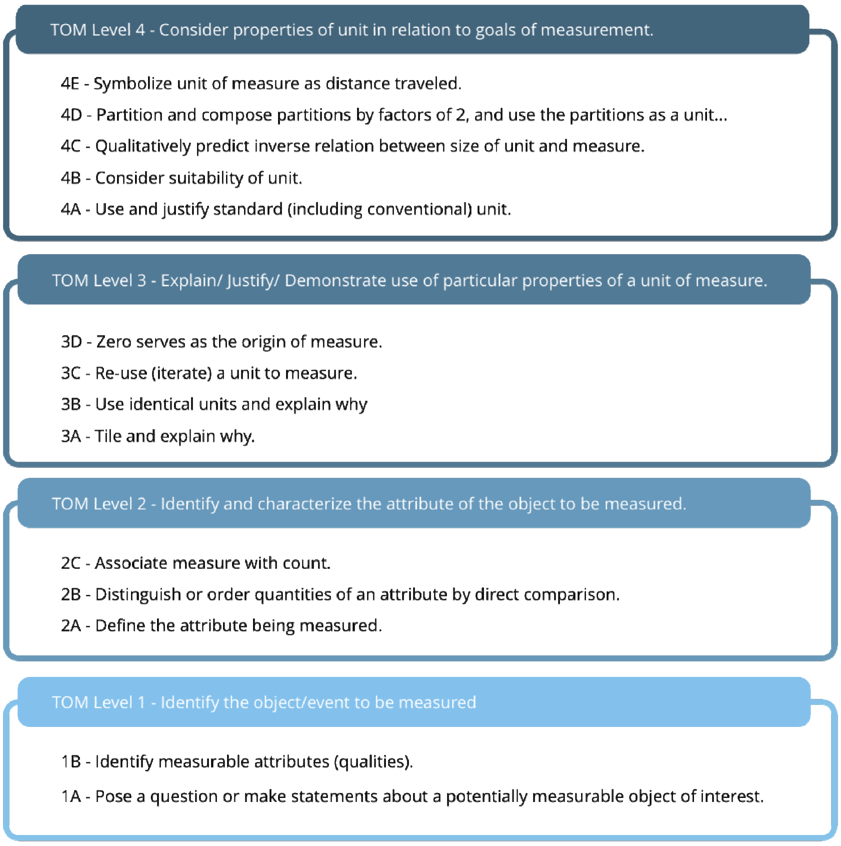


Fig. 2. Structure of the BEAR Assessment System Software

3. A sample application: The Analysis and Data Modeling Project

The Analysis and Data Modeling (ADM) project is a middle school mathematics curriculum designed by researchers at Vanderbilt University (led by Rich Lehrer) in collaboration with the educational assessment team at the BEAR Center at Berkeley [16]. The assessments being designed and tested are intended to support the development of middle school students' understandings of concepts and practices of data modeling. Data modeling refers to the invention and revision of models of chance to describe the variability inherent in particular science processes. The construction of the assessments involved analysis of core concepts and practices of data modeling that were useful to students for learning, to teachers for developing their instructional practices, and to assessment developers for developing assessments. In all, seven constructs span the range of the curriculum. The seven constructs are: Data Display, Meta-Representational Competence, Conceptions of Statistics, Chance, Modeling Variability, Theory of Measurement, and Informal Inference.

Defining the construct to be measured involves specifying a logical sequence of increasing evidence of ability, where later levels build on previous accomplishments. Fig. 3 shows an example from the ADM project for the Theory of Measurement construct. Each level of the construct corresponds to a specific level of student understanding. Each construct map characterizes the expected levels of student understanding and the main learning performances associated with each level.



TOM Level 4 - Consider properties of unit in relation to goals of measurement.

4E - Symbolize unit of measure as distance traveled.
4D - Partition and compose partitions by factors of 2, and use the partitions as a unit...
4C - Qualitatively predict inverse relation between size of unit and measure.
4B - Consider suitability of unit.
4A - Use and justify standard (including conventional) unit.

TOM Level 3 - Explain/ Justify/ Demonstrate use of particular properties of a unit of measure.

3D - Zero serves as the origin of measure.
3C - Re-use (iterate) a unit to measure.
3B - Use identical units and explain why
3A - Tile and explain why.

TOM Level 2 - Identify and characterize the attribute of the object to be measured.

2C - Associate measure with count.
2B - Distinguish or order quantities of an attribute by direct comparison.
2A - Define the attribute being measured.

TOM Level 1 - Identify the object/event to be measured

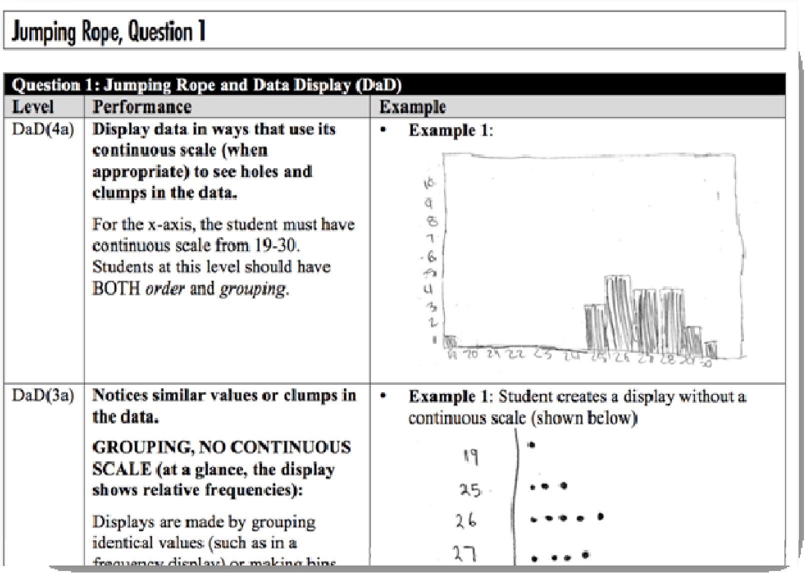
1B - Identify measurable attributes (qualities).
1A - Pose a question or make statements about a potentially measurable object of interest.

Fig. 3. Levels in the Theory of Measurement construct

Assessment items are developed with the intention of targeting specific construct levels (Fig. 4). For each item, an exemplar, or scoring guide, describes the learning performances and sample responses. Each item is explicitly related to one or more constructs through the scoring guide. Each level of a scoring guide has a description, taken directly from the description of the level of the construct, a rationale, which outlines in detail how responses to this item can correspond to that level of the construct, and one or more examples of student responses at this level. Whenever possible, these responses are taken from actual student productions.

4. Future directions

The formative assessment application BASS is built on the idea of measurement scales that stand for generalized learning structures. As the stability of these structures becomes better understood in the context of education practice, they will naturally be linked together in common systems, not unlike the metrological systems of weights and measures we take for granted in commerce and the natural sciences [17-18]. Common languages like these will eventually impact human, institutional, and information resources in ways that will make educators and employers better able to work together to identify and meet human resource and a wide variety of other needs.



Jumping Rope, Question 1

Question 1: Jumping Rope and Data Display (DaD)		
Level	Performance	Example
DaD(4a)	**Display data in ways that use its continuous scale (when appropriate) to see holes and clumps in the data.** For the x-axis, the student must have continuous scale from 19-30. Students at this level should have BOTH *order* and *grouping*.	• **Example 1:**
DaD(3a)	**Notices similar values or clumps in the data.** **GROUPING, NO CONTINUOUS SCALE (at a glance, the display shows relative frequencies):** Displays are made by grouping identical values (such as in a frequency display) or making bins	• **Example 1:** Student creates a display without a continuous scale (shown below)

Fig. 4. Sample ADM items from the Data Display construct

5. References

1. Mari L 2007 *Measurement* **40** 233-242
2. National Institute for Standards and Technology 2009, 20 July *Outputs and outcomes of NIST laboratory research* http://www.nist.gov/director/planning/studies.cfm
3. Joint Committee for Guides in Metrology 2008 *International vocabulary of metrology: Basic and general concepts and associated terms, 3rd ed* (Sevres, France: International Bureau of Weights and Measures)
4. Rasch G 1960 *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980) (Copenhagen, Denmark: Danmarks Paedogogiske Institut)
5. Dawson T L 2004 *Journal of Adult Development* **11** 71-85
6. Andrich D and Styles I M 1998 *Psychological Methods* **3** 454-469
7. Pelton T and Bunderson V 2003 *Journal of Applied Measurement* **4** 269-281
8. Mintrop H and Sunderman G L 2009 *Educational Researcher* **38** 353-364
9. Wilson M 2004 *Towards coherence between classroom assessment and accountability (National Society for the Study of Education Yearbooks vol 103, part II)* (Chicago: University of Chicago Press)
10. De Boeck P and Wilson M (Eds) 2004 *Explanatory item response models* (New York: Springer-Verlag)

11. Stenner A J, Fisher W P Jr, Stone M H and Burdick D S 2013 *Frontiers in Psychology: Quantitative Psychology and Measurement* **4** doi: 10.3389/fpsyg.2013.00536
12. Wilson M 2005 *Constructing measures: An item response modeling approach* (Mahwah, New Jersey: Lawrence Erlbaum)
13. Wilson M 2004 Towards coherence between classroom assessment and accountability (National Society for the Study of Education Yearbooks vol 103, part II) (Chicago: University of Chicago Press)
14. Wilson M and Sloane K 2000 *Applied Measurement in Education* **13** 181-208
15. Wilson M 2009 *Journal of Research in Science Teaching* **46** 716-730
16. Lehrer R, Kim M J, Ayers E and Wilson M 2014 Toward establishing a learning progression to support the development of statistical reasoning *Learning over time: Learning trajectories in mathematics education* ed A Maloney, J Confrey and K Nguyen (Charlotte: Information Age Publishers) in press
17. Wilson M and Draney K 2002 A technique for setting standards and maintaining them over time *Measurement and multivariate analysis* ed S Nishisato, Y Baba, H Bozdogan and K Kanefugi (Tokyo: Springer-Verlag) pp 325-332
18. Fisher W P Jr 2009 *Measurement* **42** 1278-1287

Acknowledgments
This work was supported by a grant from the US Institute for Education Sciences (award R305A120217). The development of the constructs and classroom implementation of the ADM curriculum is being led by Richard Lehrer and Ryan Seth Jones (Vanderbilt University). Technical staff contributing to the software portion of the project include Ana Maria Albornoz Reitze, Shazi Khan, Daniel Stanfield, Sevan Tutunciyan, Richard Vorp and Mike Kendall (all with the University of California, Berkeley). Additional recent work on psychometric analyses was contributed by Perman Gochyyev and Shonte Stephenson (University of California, Berkeley).