Metrologically coherent assessment for learning: what, why and how

Emily Pey-Tee Oon¹, U Hoi-Ka¹, W P Fisher Jr²

¹ Faculty of Education, University of Macau, Taipa, Macau SAR, China
² BEAR Center, Graduate School of Education, University of California, Berkeley, Berkeley California 94702 USA

E-mail: peyteoon@umac.mo

Abstract. Assessment for learning aims to improve learning by feeding back information on the attainment of intended learning objectives to students and teachers. In this form of assessment, the student acts as a measuring instrument and learning attainment is the quantity to be measured. Current practices in classroom assessment, unfortunately, often are not grounded and have been focused on total correct scores without reference to the evidence hidden within them concerning the attainment of learning objectives. Our goal in this study is to clarify what assessment for learning is, why it is important, and how it works in a metrologically coherent context of equated assessments administered from calibrated item banks. The main focus lies in how item content can be framed according to intended learning outcomes that are meaningfully interpreted using a construct map. An example involves empirical datasets of 283 secondary three students from two formative tests designed by two science teachers. Coherent formative assessment of this kind applies a scientific model that belongs to the same class metrologists consider as defining measurement. The results illuminate the potential for establishing item banks that enable teachers to more efficiently and effectively assess and improve the attainment of learning objectives across different cohorts of students in a common metric.

1. Introduction

1.1. What assessment for learning is
Teaching practices in exam-oriented classrooms continue to be counterproductively driven by scores decades after improved methods have been made available [1]. Higher scores seem desirable, since a higher performance rewards both teachers and students. A lower score, in contrast, is seen as a sign of failure for both teachers and students. This kind of assessment, termed as assessment of learning focuses not on the attainment and promotion of learning outcomes but on students’ abilities to perform on tests after instruction is completed.

Assessment for learning (formative assessment) differs, then, in that it occurs in classrooms during the course of instruction, and sets out to foster learning and improved outcomes. A large and significant body of published research studies have decisively established that formative assessments can move learning forward in highly productive ways not otherwise obtainable [2]. The promising potentials of formative practices, however, have not been fruitfully realized. Not only has formative assessment not been as widely implemented as one would expect given its proven value, it is often not understood as a source of information relevant to move learning forward, and even worse, it is often seen as a means of
punitive evaluating teachers and students [3], especially in cultures where formative assessments are often conducted in a more formal manner [4].

1.2. Why assessment for learning is useful, and its problems
The attainment of learning objectives needs to be regularly assessed so that instruction can be planned to extend students’ growth and to identify current levels of learning [3]. Assessment for learning may be structured formally or informally [4]. Informal formative assessments [5] are mainly verbal, as when teachers pose questions and students answer, allowing teachers to obtain multiple rapid indications about students’ current states of understanding. In some cultures, education systems rely heavily on exam-based assessment of learning, feedback is sought in writing to obtain formal evidence about learning attainment. But instead of looking into the responses of students for evidence of what the student is prepared to learn next, total correct scores tend to focus attention only on an overall level of understanding. Summed total correct scores do not provide teachers with information useful to improve their ongoing instructional planning, and neither does it provide students with information useful to inform their learning processes relative to their objectives. It becomes instead a tool for ranking students, distorting the educational process. As a consequence, formative assessments implemented in this context, where a needed shift in values and practices has not been put into effect, have tended to become an additional interim assessment of learning. There is still much to be done with assessment of this kind are to be acted as agent for learning [1, 6].

2. How to plan and evaluate assessment for learning

2.1. Framing item content according to intended learning outcomes
Bloom’s cognitive taxonomy [7] is often used as a guide in the framing of intended learning outcomes for knowledge and understanding development in education where construction of knowledge is conceptually interpreted in a progressive scale: (a) remember, (b) understand, (c) apply, (d) analyse, (e) evaluate, and (f) create.

Theoretically, test item ought be written at a given level of cognitive challenge tapping an intended learning outcome. Taking an intended learning outcome (LO) on net forces for an example:

(a) A lower cognitive LO: students are able to recall that net force of a stationary object is zero
(b) A higher cognitive LO: students are able to explain that an object is stationary because either the net force acting on it is zero or no force is acting on it.

These contrasting assessment focal points are both borne out from an intended LO in which students are expected to understand that an object is not moving either because there are no forces acting on it or because the net force acting on it is zero. A correct response to an item with lower difficulty level implies that student has attained understanding that there is either no force or a net force of zero for a non-motion object. A correct response on a second item assessing higher cognitive understanding implies that net force applied to an object is actually proportional to its resulting acceleration [8]. Students who can answer the higher cognitive LO correctly most likely can answer the lower cognitive LO correctly. However, students who can answer the lower cognitive LO correctly may not be able to answer the higher cognitive LO correctly. Content difficulty for the higher LO is theoretically higher than the lower LO. Not basing expectations on personal or subjective judgements requires mapping content difficulty against cognitive complexities of intended learning outcomes relative to both theoretical and measurement analysis justifications.

2.2. Justification of theoretical and measurement analyses of intended learning outcomes
Items framed to assess higher and more complex cognitive LOs are likely to be estimated as being more difficult as compared to other items framed to assess lower and less complex cognitive LOs. Measurement models capable of framing and testing this kind of hypothesis are useful in quantifying learning outcomes explicitly as repetitions of an invariant unit. These model definitions and experimental tests are recognized by metrologists [9-10] as scientific approaches to measurement:
which says that the log-odds of observed success for student \( n \) on item \( i \) at partial credit score \( j \) is equal to the difference between the estimate \( B \) of person \( n \)'s ability and the difficulty estimate \( D \) of item \( i \) at rating \( j \) relative to the next lower rating of \( j \) [11-13]. Estimates expressed in these log-odds units (logits) are mathematically and experimentally proven to exhibit linear, interval, and additive properties [9-11].

Interpretation of these units is informed by construct theory, an explanatory model predicting variation in terms of causal relations [14-17]. For example, an item framed to assess content of middle difficulty ought to arrive at calibration lower than another item framed to assess content of high difficulty. If the measurement analysis results are not aligned with what has been expected theoretically, reasons to explain “why” and revisions of item content are in order. Rounds of iterations through this process to achieve stable relationships between theory and item calibrations are inevitable.

2.3. Formation of item strands based on theoretical construct map
Repeated patterns of calibrations emerging spontaneously across samples of students and across assessments indicate when construct definitions have achieved a stable balance between theoretical and measurement expectations. Theoretical terms can be communicated as strands in a construct map [14-15] outlining meaningfully varying amounts of the trait measured. Construct maps can be used to guide the writing of assessment items and to track learning progressions [15-17] from novice to expert understanding over time for each individual student.

3. An example
Three items were written from theory to assess different difficulty levels. Item 1 examines students’ understandings at the lowest difficulty level, and item 3, the highest level (Table 1). The empirical scaling results, however, do not always align with theory. Two estimates (and uncertainties) of the difficulty of item 1, at 0.02 (0.07) and 0.03 (0.07) logits respectively, from two separate calibrations, are higher than estimates from the same samples for item 2, at 0.53 (0.1) and 0.46 (0.01) logits, even though the opposite was expected.

Item 2 was designed to examine if students are able to apply what they have learned about the food pyramid in general to a problem of actual daily relevance, whereas item 1 should elicit only the easier recall task. The measurement result has promoted the teacher who designed item 2 to think that this item is not at the level of applying existing understanding but instead involves simply recalling food categories according to different pyramid levels.

Theoretical and measurement results comparative alignments of this kind aim at defining a coherent assessment objective that serves to move student learning forward by identifying what is known, what is not known yet, where the student stands relative to the LO. Once stable relationships are established through multiple repeated measures, a provisionally definitive construct map can be built to inform interpretations of the trait being measured (Table 1). For an example, level 1 in the construct map outlines the minimum amount of knowledge students have achieved, and level 3 outlines the highest capacity of knowledge achieved by students with regards to the learning objectives. From here, teachers can feed back the information to students on what they have known and, most importantly, what they do not know yet. A response at a partial credit level may possibly indicate the student is hindered by misconceptions that must be addressed to move learning forward. Conversely, information on student performance can be fed back to teachers, communicating on the effectiveness of their instruction. Learning activities can be modified if most students fail to achieve the learning objectives. Having results measured in a unit of measurement comparable across classrooms provides additional opportunities for teachers to learn from each other and from their own historical data.
Table 1. Construct map pertains to learning objectives with their test items.

| Learning Objective | Bloom's Taxonomy of Learning | Construct Map |
|--------------------|------------------------------|---------------|
| Students are able to tell the categories of food ingredients in each layer of food pyramid. | Remember (Item 1) | From the following categories, which layer in food pyramid does the food belong to? | 0 - Cannot correctly describe the classification of each layer in the food pyramid. |
| | Apply (Item 2) | Mango, broccoli, butter, biscuits, fresh milk, cheese, granola, chocolate | 1 - Can clearly understand the classification of the layers in the food pyramid but cannot put all kinds of food into the corresponding categories. |
| | Analysis (Item 3) | William studied the food pyramid shown on the left. He then interpreted it as shown on the right, as below. | 2 - Can clearly understand the classification of the layers in the food pyramid and put various foods into each corresponding categories. |
| | | Explain why William’s idea about each level is not completely correct. | 3 - Students can choose the right food from the food pyramid according to the needs of individuals. |
| | | (a) Fat products | |
| | | (b) Protein products | |
| | | (c) Dietary fibre products | |
| | | (d) Carbohydrate products | |
| | | Answer: | |
| | | A: butter, chocolate; B: fresh milk, cheese; C: Mango, Broccoli; D: biscuits, wheat husks | |

4. Discussion
Student-focused assessment for learning reveals misconceptions useful for facilitating student progress toward learning objectives, where teachers are able to provide meaningful feedback to students in closing learning gaps. Teachers are similarly provided feedback on their own performance, so they can also evaluate and improve their instruction.

The example given demonstrates some of the student- and instruction-focused fundamental requirements for meaningful assessment for learning. This framework sets the stage for establishing item banks measuring the attainment of related learning objectives. Items in a bank can be created to
provide reliable assessments across difficulty levels associated with specific learning objectives, and that are psychometrically validated as repeatedly and reproducibly measuring a construct. With tools like that available, teachers can select, apply, construct, and calibrate items for use in classroom assessments, adding to and working from the banks of questions.

5. Conclusion
Establishment of item banking of this kind will set stage to combine assessments for learning (formative assessment) and of learning (summative assessment) in a developmental coherence scale, as below:

Figure. 1. Developmental, horizontal, and vertical coherence (TP=Time Point; A1.1-A2.4 are content domains) [18].

Summative evidence can be innovatively produced from accumulated formative assessments over time, instantiating the developmental coherence shown in figure 1. To achieve this purpose, formative evidence has to be collected in a reliable and valid form, meaning measures must be estimated and reported. This requires formative assessments and feedback to be structured in formal ways documenting progress [4]. But in this context, students’ growth in understanding can be charted over time, allowing meaningful interpretation to move learning forward and to inform instruction. Capacities for taking missing data into account, adaptively administering items, and resiliently modifying LOs as educational needs and circumstances change suggest a productive future for this approach.

References
[1] Black P and Wiliam D 2018 Assessment in Education: Principles, Policy & Practice 25 1-25
[2] Hattie J 2008 Visible Learning (New York: Routledge)
[3] Heritage M 2007 Phi Delta Kappan 140-5
[4] Dolin J, Black P, Harlen W and Tiberghien A 2018 Transforming Assessment eds J Dolin, R Evans (Switzerland: Springer Nature) pp 53-80
[5] Duckor B 2014 Educ. Leadership 71 28-32
[6] Yan C 2015 Improving Schools 18 5-19
[7] Anderson L W and Krathwohl D R 2001 A Taxonomy for Learning, Teaching, and Assessing (New York: Longman)
[8] Alonzo A and Steedle J 2009 Sci. Educ. 93 389–421
[9] Mari L and Wilson M 2014 Measurement 51 315-27
[10] Pendrill L and Fisher W P Jr 2015 Measurement 71 46-55
[11] Rasch G 1980 Probabilistic Models (Chicago: University of Chicago Press)
[12] Andrich D 1978 Psychometrika 43 561-73
[13] Masters G N and Wright B D 1997 In W J van der Linden and R K Hambleton Eds, Handbook of Modern Item Response Theory pp 101-21 (New York: Springer-Verlag)
[14] Roberts L, Wilson M and Draney, K 1997, June The SEPUP Assessment System: An Overview. BEAR Report Series, SA-97-1 (Berkeley: University of California)
[15] Wilson M 2005 Constructing Measures (Mahwah, New Jersey: Lawrence Erlbaum)
[16] Wilson M R 2009 J. Res. Sci. Teach. 46 716-30
[17] Wilson M R and Sloane K 2000 Appl. Meas. Educ. 13 181-208
[18] Fisher W P Jr, Oon E P-T and Benson S 2019 Measurement in review