The Assessment Evaluation Rubric: Promoting learning and learner -centered teaching in higher education

Rochelle E. Tractenberg

Rochelle Tractenberg
Building D, Suite 207
Georgetown University Medical Center
4000 Reservoir Rd. NW
Washington, DC 20057 USA
rochelle.tractenberg@gmail.com

Tractenberg, RE. (2020, 20 April). The Assessment Evaluation Rubric: Promoting learning and learner -centered teaching in higher education. Published in the *Open Archive of the Social Sciences (SocArXiv)*, DOI 10.31235/osf.io/bvwhn

Shareable under a CC - by attribution - noncommercial - no derivatives 4.0 license

Running Head: Assessment Evaluation Rubric (AER)

Acknowledgement: This work supported by NSF grant: RCN-UBE DBI 182713, and/with Co-PIs Anne Rosenwald & Rochelle Tractenberg (Georgetown), Vince Buonaccorsi (Juniata College), Douglas Chalker (Washington University St. Louis), and Jason Williams (Cold Spring Harbor).

While this paper is intended to be *factual* - and the facts reported are correct - any opinions expressed are my own, and do not represent the perspectives of the National Science Foundation, Genomics Education Alliance, or Georgetown University

Abstract

It is common to create courses for the higher education context that accomplishes content-driven teaching goals, and then develop assessments (quizzes, exams) based on the target content. However, content-driven assessment can tend to support teaching- or teacher- centered instruction. Adult learning and educational psychology theories suggest that instead, assessment should be aligned with *learning*, not teaching, objectives. To support the alignment of assessments with instruction in higher education, the Assessment Evaluation Rubric (AER) was developed. The AER can be utilized to guide the development and evaluation/revision of assessments that are already used. The AER evaluates four features of an assessment: its general alignment with learning goal(s); whether the assessment is intended to/effective as formative or summative; whether some systematic approach to cognitive complexity is reflected; and whether the assessment (instructions as well as results) itself is clearly interpretable. Each dimension (alignment; utility; complexity; clarity) has four questions that can be rated as present/absent (or yes/no), or, using a three-level ordinal scale describing “present-useable”, “possibly present - needs clarification”, and “absent”. Other rating methods can also be conceptualized for the answers to the AER’s 16 questions, depending on the user’s intent. Any instructor can use the AER to evaluate their own assessments and ensure that they -or new assessments in development - will promote *learning* and *learner* centered teaching.

Introduction

Assessment is defined as “the systematic collection of information about student learning, using the time, knowledge, expertise and resources available, in order to inform decisions that affect student learning.” ([1] p. 2; see also [2] p. 4; [3] pp. 1-13). Walvoord [1] goes on to assert that “(t)he goal of assessment is information-based decision making” (p. 4). The *Standards for Educational and Psychological Testing* specify that “(t)est development is the *process* of producing a measure of some aspect of an individual’s knowledge, skills, abilities, interests, attitudes, or other characteristics...” (emphasis added; [4], p. 75). These sources agree that assessment is systematic – and its development is a process – such that the result can be utilized for measuring or decision making. The decisions that affect student learning are intimately tied to the objectives of the instruction ([3] p. 15; [5]); and Tractenberg et al. (2020) [6] discuss a five-phase model for curriculum and instructional development (based on [7]) where “assessment” is phase 4 of five, explicitly considering how closely tied to learning objectives – defined in phase 1 – and how teaching is accomplished (learning experiences, phase 2) the assessment must always be.

When assessment is integrated *within and during* the learning experience, be it course, workshop, or curriculum, it can be used to make decisions about how best to teach in order to optimize student learning. It is defined as formative assessment when decisions can most directly affect student learning. When assessment occurs after instruction ends, it is summative, providing a summary of the effectiveness of the instruction rather than information that can be leveraged to adjust teaching to optimize learning ([3] p. 19). From the student perspective, it can seem that the summative assessment of a course (e.g., final exam) or curriculum (e.g., a capstone project) is the most important one, but decisions that affect student learning are, by these endpoints, only able to affect the learning of *other*, later students. According to Walvoord (2010) [1], “(g)rades are only minimally useful for assessment; much more important are evaluations of the strengths and weaknesses of student work”. (p. 10)

“Although unguided or minimally guided instructional approaches are very popular and intuitively appealing ...these approaches ignore both the structures that constitute human cognitive architecture and evidence from empirical studies over the past half-century that consistently indicate that minimally guided instruction is less effective and less efficient than instructional approaches that place a strong emphasis on guidance of the student learning process.” [8]

The adult education, educational psychology, and cognitive science domains have long documented what works to promote learning [9]. As the quote from Kirshner et al. (2006) [8] points out, key features of human cognitive architecture, and empirical evidence about leveraging that architecture to optimize learning, have been known for decades. A key feature of a successful curriculum in higher education is that it contains opportunities for assessment that both enable trainees to demonstrate their progress in authentic, interpretable ways, and provide explicitly for the evaluation of the functioning of the curriculum itself [3, 10, 11]. Handelsman et al. (2007) [12] emphasize that the systematically-collected information from assessments provides evidence to both instructors and learners (p. 47), that each can use to improve what they

do to contribute to the learning enterprise. That is, teachers can improve instruction to promote better/more learning, and learners can improve how and what they learn ([13] p 28).

Although assessment is an essential aspect of any instructional opportunity, the first, and arguably most important, feature of a curriculum or instructional opportunity is the articulation of learning objectives [5, 14, 15]. Learning outcomes should be emphasized in any development, evaluation, and revision effort because of how central they are to all of the other decisions to be made, including conclusions about the effectiveness of a curriculum (extensively discussed in [6]). This decision-making happens most obviously on the parts of the instructor, in both formative and summative assessments, and the institution (e.g., allow into the major; define satisfactory completion of objectives of the major). However, in a curriculum that follows principles of andragogy [16], the learner is a *partner* in their education, not a vessel to be filled with knowledge; thus, decision-making by the learner is also an important consideration in the creation and evaluation of assessments for adult learners.

Handelsman et al. (2007) [12] point out that most higher education faculty do not have much/sufficient background or training in education or learning to make justifiable choices in the way they teach and assess (p. 3). Fortunately, education and cognitive scientists have documented, summarized, and synthesized relevant literatures, and converge on several straightforward principles. Both the creation and the evaluation of assessments are motivated from the consideration of four key characteristics or dimensions:

- **General alignment** with the course or curriculum (e.g., [3: Ch 4; 10, 17]);
- **Utility** of the assessment (e.g., [3: pp 72-75; 18; 19: Ch 3-4])
- **Alignment of the assessment's cognitive complexity** with the instruction the assessment is intended to capture ([5: Ch 1; 10: p. 213; 20]); and
- **Clarity of the assessment** itself [4, 21].

All assessments can be described along these four dimensions – and more to the point, when assessments do not have demonstrable alignment with the course or curriculum (dimension 1), utility (dimension 2), appropriate cognitive complexity (dimension 3), *and* clarity (dimension 4), their interpretability will be minimal/minimized. The National Research Council (2001) informally defines assessment as “a process by which educators use students’ responses to specially created or naturally occurring stimuli to draw inferences about the students’ knowledge and skills (citing Popham, 2000; [13: p. 20]). As noted earlier, formative assessment is differentiated from summative assessment by its focus on diagnosis, in order to improve both teaching (by targeting it better) and learning (by highlighting for students what specifically is deemed to be lacking [13]. The ability to take clear action on the basis of an assessment makes it *actionable*, and the National Institute for Learning Outcomes Assessment (NILOA) argue that actionable assessments are those that create evidence that can be “translated into actions to enhance student accomplishment” [22: p. 6]. Effective formative assessment is actionable for both the instructor *and the learner*, and possibly also for the institution.

Effective formative assessment requires the ability of the assessor to obtain diagnostic information about the learner and then utilize the information so acquired, in order to structure teaching and/or to promote recognition in the learner of what is missing and what specifically needs to improve. That is, the decision-making that formative assessment supports is fairly

immediate so that the instruction and learning strategies can be adapted by the two key participants in the learning enterprise, the instructor and the learner. By contrast, summative assessment is focused on the end result of the instruction for the learner- after which assessment no additional modifications will normally be made in either teaching or learning of the knowledge, skills and abilities that were summatively assessed. Summative assessment may document student achievement, and as such, is less actionable than formative assessment is/can be; by describing students after the end of an instructional opportunity, summative assessment may be less actionable for institutions than formative assessment is.

Whether summative or formative, assessment is meant to inform decision-making, and as such, the *Standards for Educational and Psychological Testing* [4] should be utilized. “Assessment is a broader term than *test*, commonly referring to a process that integrates test information with information from other sources [4: p. 2]. *Validity* refers to the interpretation of test scores – rather than to the test itself – so when test scores are interpreted to effectively support actions and decisions by the learner, the instructor, or the institution (as suggested by [22]) to improve student outcomes, the assessment that leads to those scores would be called “valid”. “Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use” [4: p. 11]. In a sense, the AER can be used by any instructor in order to describe and better understand their own arguments for making decisions about student learning, and their instruction, utilizing assessment. The AER was created based on “what works” in higher/adult education: Bloom’s Taxonomy of Cognitive Behaviors [15] to underpin a developmental model of learnable and improvable – and observable - cognitive behaviors that the course or lesson may be designed to promote; Messick’s (psychometrically) valid assessment features [17], to ensure that learning is targeted and yields observable and interpretable behaviors by students; and principles of andragogy [16], to ensure that adult learners are “getting” what the curriculum seeks to deliver. Thus, the four dimensions, and each feature in the dimensions, can either be applied to an existing assessment, or they can be used to support the construction of new assessments that will optimally provide valid information about student learning and instruction. The AER can be used for both formative and summative assessment.

Assessment Evaluation Rubric

A rubric is simply a matrix that generally describes performance in response to some direction (i.e., on a specific task). As discussed by Dawson (2015) [23], a typical rubric includes evaluative criteria and descriptions by which the ‘quality’ or ‘level’ of student responses on each of those criteria can be characterized (usually ranked from poor to excellent), and some reproducible method for assigning quality labels to student work (p. 349; see also e.g., [1, 24]). Many readers may be familiar with the use of rubrics to assess student work. However, rubrics can also be constructed and utilized to evaluate or assess *other* work as well – the AER is an example of applying the rubric construct to the evaluation of assessments. Using Dawson’s (2015) [23] rubric design element vocabulary, five of the 14 rubric design elements are most relevant for the AER:

- Task specificity
- Quality definitions

- Secrecy
- Scoring strategies
- Quality levels

The object of the AER is *task specific* where the task is the creation of an assessment by the instructor. The AER is applied to a given assessment, or used to ensure that a new assessment is created to have the four key dimensions that will support valid uses of the assessment. Although the AER is task-specific, it has implied, but not explicit, *quality definitions* in the sense that there are four key dimensions of assessments that optimize the information that an assessment can provide, but the assessment developer or evaluator brings their own value judgments about the relevance of the dimensions to their own decision making. For example, in a mentoring situation, the mentor would possibly use more elaborate quality definitions to promote the revision and improvement of an assessment (so it becomes more consistent with all features in the AER), whereas for a group that is choosing assessments for inclusion or recommendation, the quality definitions may be used to make decisions about which assessments to keep/recommend and which to eliminate. The AER is never *secret*, and it and the result of applying it to an assessment are meant to be shared widely, as it represents a synthesis of empirical work on education and the validity of actionable decisions in higher education (and the author is a proponent of always sharing what is being assessed with those being assessed!). Like the quality definitions, *scoring strategies* to be used with the AER are up to the user; ultimately, the purpose of *creating* the AER is to allow all instructors to achieve “full marks” on all four dimensions for every assessment that they utilize or create. Finally, because it describes general educational, psychological, and psychometric attributes of assessment, the *quality levels* are open to selection by the user of the AER, although as will become apparent with the features of each dimension, each can be rated “present”/“absent”, and if one “point” is assigned for every present feature, the maximum “total score” on the AER would be 16.

The four-dimensional rubric was created in order to support, and carry out, the systematic evaluation of assessments, and to promote the creation of new assessments that are optimized to support learning outcomes in a valid way. The four dimensions, listed above, pertain in all educational contexts but this rubric and the discussion below was developed specifically for higher education.

Each of the four dimensions has four questions, listed and discussed below.

Dimension 1: General alignment with the course or curriculum

1. A learning goal was articulated (that this assessment is intended to assess).
2. The assessment is aligned with the learning goal(s).
3. The teaching is supportive of what is being assessed.
4. The assessment is aligned with the Bloom’s level complexity of thinking about the content that the learner should have gained from the instruction.

The first of the four dimensions to the Assessment Evaluation Rubric is General Alignment. This dimension is intended to capture the relevance of the assessment for the instruction (see e.g., [3, 10, 17]). An assessment should pass (score 1) on all of these features of the alignment dimension, or else the rest of the dimensions will not be meaningfully rated for that assessment.

The first feature required to determine the alignment of an assessment is that a learning goal or outcome was articulated - in order to evaluate the alignment of any assessment with the learning it is intended to assess, the learning (NB: *not teaching*) goal must have been articulated. If not, then the assessment cannot work – it will not be valid for the assessment of learning (although it could possibly be informative about teaching). All four elements of the alignment dimension that are assessed with this dimension of the rubric are tied to Messick’s three features of valid assessment [17]:

- What are the knowledge, skills, and abilities (KSAs) the curriculum should lead to?
- What actions/behaviours by the students will reveal these KSAs?
- What tasks will elicit these specific actions or behaviours (that reveal KSAs)?

These questions are also embedded in Wiggins & McTighe’s Backwards Design (2005) [25], and with other well-established cognitive psychological principles.

Whatever tasks students are asked to carry out in the assessment (i.e., Messick #3) should have some relationship with *observable verbs* - because to accomplish Messick #2 the actions that students do to demonstrate they have learned what was intended must be observable (and, although it is by no means the only source, verbs from Bloom’s Taxonomy [15] work for this; see also [26]). If you keep the three Messick questions in mind as you develop any assessment, then it will be more likely to achieve/the assessment will satisfy all of the features in the alignment dimension.

Because this is a rubric, there is a lot of flexibility in exactly how, and with what content, the assessment can exhibit these features with respect to the target learning materials. There may or may not be interest among the users of the AER in ensuring or documenting the alignment of the assessment with national (or other) compendia of content, concepts, or competencies, although these can be combined with/utilized as the learning objectives in element #1, the identification of specific learning objectives about which the assessment is intended to provide evidence.

Importantly, it may not be feasible to expect that all evaluators (users of the AER) will know what is the “correct” answer to AER items – particularly on the alignment dimension. However, when multiple evaluators, including the individual who created the assessment, apply the AER to the same assessment(s), discussions can be initiated about the course or teaching and learning objectives, and how the assessments support both teaching and learning in actionable ways. An understanding of the alignment of the assessment to learning objectives is essential for an interpretable, valid, assessment.

Dimension 2. Utility of the assessment

1. The assessment is documented as intended to be formative (for the instructor, learner, or both).
2. The assessment is (also) intended to be summative.
3. The results of the assessment are informative about the learning that the instructional opportunity was intended to provide.

4. Assessment item structure is correct (e.g., common errors of multiple-choice item construction are avoided; stems do not suggest the answer; the correctness of a multi-part item is not contingent on subparts that are at diverse levels of complexity).

The second dimension on the AER is utility. As noted, a valid assessment is one with strong evidence that supports interpretation of its scores (or summary) ([4]; see also [3, 18, 19]). Formative assessment is intended to provide input/identify what is and what is not being learned/acquired. Summative assessment happens after the learning ends, and summarizes the learning that took place. The same assessment cannot coherently be used for the two purposes simultaneously. Raters using the Assessment Evaluation Rubric would need to rate this feature, but what to do if an assessment is (mis)labeled as both formative and summative, or just mislabeled as formative when it is better/more clearly summative, or other types of problems that might be observed in a given assessment being evaluated on this dimension, is not clear. One thing to consider is, if an assessment is evaluated on this dimension and gets 0/4 “points”, would the assessment be rejected, returned for revision, or simply not be approved for inclusion in an otherwise badge- or approval-appropriate stamp? The features in this dimension may be easier to correct (if in need of revision, or if they do not exist) than for Dimension 1 (alignment). That is, if a single assessment is intended to be both summative and formative, then the assessor can be notified of this and they can simply choose which use they prefer. By contrast, more work would be needed in order to revise/correct any of the elements in dimension 1 so that they ‘earn’ each point. However, assessment item structure (question 4 in the Utility dimension) is a well-known underminer of what could otherwise be effective assessment – because good multiple-choice items are difficult to write, particularly at a specific cognitive complexity level [27]. Moreover, if question 3 is rated “no”, it would suggest that the assessment cannot be characterized as a useful assessment – so it would need to be changed sufficiently so that the results are in fact informative (i.e., it would need to be made into a valid assessment), and that could be very time consuming and effortful.

Importantly, if ‘scores’ on dimension 1 (alignment) are low (or all 0), the utility of the assessment will similarly be low or lacking. If dimension 1 ratings are low and somehow, ratings on dimension 2 are higher, this will point to misinterpretations of the AER dimensions rather than a useful instrument that is not aligned with instructional purposes/objectives. Once alignment and utility features are fully present for an assessment, then a deeper examination of the role of cognitive complexity in the structure of, as well as information gleaned from, the assessment can fruitfully be examined.

Dimension 3. Alignment of the assessment’s Bloom’s (or similar) level complexity with the instruction the assessment is intended to capture

1. Bloom’s taxonomy (or equivalent) is used (correctly and appropriately, i.e. not just for memorization, unless that is exactly the target of instruction).
2. Rationale for using specific levels, or manipulating, the cognitive complexity in items is clear/coherent.
3. There is attention given to the cognitive complexity of the items.
4. The assessment was intended to provide the student with actionable evidence of what to do (better next time/to take advantage of the learning that was achieved).

Cognitive Complexity is a dimension of both teaching and assessment that focus all attention on the learning and the learner. Specifically, consideration of the cognitive complexity that an assessment requires should include determining both a) what levels of complexity a learner *can rise to* as a result of a specific learning experience; and b) what levels of complexity are required to respond to/answer questions on the assessment (see [27, 28]). If the instruction is geared towards critical thinking, which requires more complex cognitive capability than recall, or the application of rules or execution of an analysis following rules or a set of specific instructions (recipe), then the assessment should not be limited to multiple-choice questions (which typically engage Bloom’s levels 1 (recall) and 2 (understanding), see [27]). If an instructor articulates a learning outcome (goal) that includes critical thinking, then the learning to be demonstrated on the assessment should feature *that kind* of cognitive complexity. If the assessment includes critical thinking, then instruction should also focus on ensuring that the abilities to think critically can be learned, with feedback, throughout the instruction – implementing Messick’s three key questions [17]. Thus, ensuring an assessment is aligned on this dimension can promote an emphasis on learnING, and the learnER, rather than on teaching and the teacher, throughout instruction.

While “active learning” is an important consideration for how teaching is delivered (e.g., [12]), effective higher education can also ensure that learners specifically develop increasingly complex cognitive abilities (e.g., [29]). Achieving the complexity features of an assessment may be facilitated when dimension 1 (alignment) is fully satisfied (i.e., all alignment features rated 1), but there may be interest in prioritizing alignment and leaving complexity for “more sophisticated” assessment development. “Actionable evidence” means that, on the basis of the observed response on any assessment (or how that was graded), the learner (or the instructor) can formulate a plan for how to proceed/continue to learn - preferably one that is more sophisticated than “I need to study harder”. Instructors can utilize Bloom’s taxonomy ([15] see also [27]) or any relevant framework from the compendium compiled by Moseley et al. (2005) [26]. Satisfying the complexity dimension can be simplified when learning outcomes (the focus of dimension 1, alignment) are consistent with criteria such as the National Institute for Learning Outcomes Assessment (NILOA, 2016) [22] criteria. However, ensuring there is consistency and alignment between the instruction the assessment is intended to capture - e.g., not teaching with memorization in mind and then asking for critical essays in assessment – is also facilitated when learning outcomes are clearly stated, and both teaching and assessment are well-aligned with these outcomes (see [5: Ch 1; 10: p. 213; 20]).

There are many cognitive complexity taxonomies (see [26]). In the US, Bloom’s [15] is the most common/familiar one, but even so, it is not often used well or fully. When using a taxonomy for developing an assessment (or items/questions on an assessment), similar verbs should also be featured in the articulated learning outcomes. In their synthesis of the literature around curriculum and instructional development, Tractenberg et al. (2020) [6] point out that effective higher education is recognizable as instruction where learning objectives drive all decisions, including those relating to the development of assessment. Based on our experience, or simplicity of scoring decisions, instructors may believe we “need to have 10 items on a quiz” or “need at least one essay answer” on any test or quiz. If the items are all simple factual questions (low Bloom’s level), ten of these might not be necessary. If the learning objective is to “encourage critical thinking”, asking students to demonstrate their memorization of ten facts

cannot support that objective – but, critical thinking cannot be assessed easily with multiple choice exams. Essays, even short answer, can be hard to grade at scale and especially, grade consistently across responses. However, these are not coherent reasons for why ten multiple-choice items would be included in any assessment. In addition to considering the cognitive complexity of your learning objectives as you construct or select the items for your assessment, it is also important to keep in mind how you will consistently grade or rate student work. (Rubrics are an interesting option for essays and other narrative response-based essays; see [24]).

Finally, assessments can sometimes include interstitial messaging like, “this section tests your ability to knit with large needles”. While it is helpful to orient the learner towards the purpose of sections of the assessment, if items vary widely in the cognitive complexity they require, and particularly if this does not seem orderly (e.g., moving from simpler to more complex), then the rationale for that complexity – even if systematic – might not be coherent for the learner, undermining the good instinct to provide that orientation. In fact, coherence in the content of the assessment is the focus of the fourth AER dimension, clarity.

Dimension 4. Clarity of the assessment

1. Students would understand clearly what they are required to do “for a good grade”.
2. Instructor would be able to interpret the summary of student work on the assessment (total score or label <pass/fail>).
3. Items on the assessment are exchangeable (so that a total score, if that is computed, is interpretable). Getting one item wrong provides the same information about student learning as any other item.
4. The items on the assessment are there for a clear reason– and appropriate for the learner’s level of experience, in terms of required cognitive complexity.

Once all the other three dimensions have been rated “present and acceptable”, then the “final version” of any assessment should be evaluated for its clarity (see e.g., [4, 21]. It is difficult to keep clarity in mind while addressing the other three dimensions of the AER, so this dimension appears last because it is the last one to finalize. However, the clarity of the item(s) and of the signal that the assessment provides to the instructor and the learner are essential to effective assessment. Modifications to improve clarity can include reordering items, integrating examples or better/clearer instructions, and other, similar helpful information that inform those being assessed about what exactly they need to do to demonstrate what or how well they have learned. Leveraging Messick’s three questions [17] in the creation of assessment items can strengthen the clarity of any test or quiz; and the construction of a rubric can be similarly helpful to the developer of the assessment, as well as those whose work is being assessed. In addition to utilizing Messick’s three questions for item development, an approach to ensuring that the test will generate an interpretable signal to the instructor about the learner’s achievement is a *table of test specifications* or test blueprint [30, 31, 32]. At its simplest, a table of test specifications is a table that specifies the features of your test. It lists the contents, distribution of questions (number, topics, cognitive complexity, and/or item type).

Two example tables appear below. The first table shows the representation of five topics on a single test:

30%	12%	40%	9%	9%	100%
TOPIC1	TOPIC2	TOPIC3	TOPIC4	TOPIC5	*TEST*

The second table shows the distribution of item types within a single test:

30%	12%	40%	9%	9%	100%
essay	short answer	Multiple Choice Question (MCQ)	fill in	matching	*TEST*

It is straightforward to combine these two features, item type and topic, into a single table that describes one test (where row totals are 100% of items on that topic):

	essay	short ans	MCQ	fill in	matching
TOPIC 1	30%	19%	50%	1%	0%
TOPIC 2	30%	11%	20%	5%	34%
TOPIC 3	20%	20%	40%	10%	10%

A table of test specifications – for a given assessment or for all of the assessments in the course – can also be constructed to ensure that the assessments are both aligned with the learning objectives for the course, and also that they include item types that require the targeted level of cognitive complexity:

30%	12%	40%	9%	9%
essay	short answer	MCQ	fill in	matching
Objectives 1, 4	Objective 2	Objectives 1-4	Objective 3	Objective 4

However, in keeping with AER dimension 3, it is perhaps more important to consider how cognitively complex (in Bloom's, or B, levels) the assessment of learning on each topic will be in the assessment, as describe in the table below showing the distribution of items across Bloom's levels (B1 – B4) on a two-topic assessment:

	Reiteration (B1)	Summarization (B2)	Illustration (B3)	Prediction (B4)	ROW totals
TOPIC 1	30%	19%	0	1%	50%
TOPIC 2	10%	10%	30%	0	50%
Column Totals:	40%	29%	30%	1%	100%

The table above shows that the assessment covers Topics 1 and 2 equally, and that the majority of assessment questions require Bloom’s levels 1-3 (B1, B2, B3). In this example table, “illustration (Bloom’s level 3, B3) and “prediction” (Bloom’s level 4, B4) are included in the table – suggesting that the assessor is actually interested in whether or not students can demonstrate this level of complexity in their responses. However, B3 is only tapped by items on Topic 2 – which might make sense if Topic 1 is more foundational knowledge and Topic 2 is focused on utilizing that knowledge; but since there is only 1% of items (i.e., one of 100 items) at B4 – and this item assesses Topic 1, this table shows that there might be a mismatch between the instructor’s learning goals and how they are being assessed. This example table therefore shows an exceptionally important use for tables of test specifications: they can help instructors (and assessors) ensure that they are both teaching and assessing, across topics, at cognitive complexity levels that are consistent with their learning objectives [33, 34].

The exchangeability of items is an essential feature for using a “total score” - although this is almost never considered or discussed when people are developing or being taught to develop assessments. Exchangeable items are those where, if two people each miss two out of 10 items, then their equivalent scores of 8 points are identical in the information the scores provide. That is, it doesn’t matter which two items of the 10 are incorrect, those two scores of 8 are identically informative and the test takers who earn 8 points are also indistinguishable in their level of learning. It is rare that tests and quizzes have truly exchangeable items on them, and yet we tend to treat all scores of the same level or value as identical (everyone scoring 70-75 points out of 100 gets the same grade, meaning that it does not matter which 25-30 points were incorrect). Using the table above, you might characterize all of the items at Bloom’s levels 1-2 as exchangeable, and then generate a “total score” for those items, with separate “subtotals” for items at level B3 and at level B4. In this table, B4 is only on 1% of the items (1 of 100 items) – which is typical in multiple-choice question exams (as outlined by [27]). The table above shows an assessment that does not assess students’ abilities to reason – nor are respondents employing any higher cognitive complexity beyond memorization, summarization, and illustration. If that is the learning objective – which will have been documented in AER dimension 1, then this table shows that the instruction and assessment will achieve that objective. However, if the abilities to remember, summarize, and illustrate are NOT the learning objective, then a table like this one shows clearly that the summary of learner performance on an assessment that matches this table will not be useful to the instructor.

An instructor can use a table of test specifications for any assessment, to accomplish (or plan for, or evaluate the extent to which the assessment):

- Identifies areas to emphasize in class or on homework (to adequately prepare students for exams);
- Ensure a test or assessment is covering those topics/skills covered in class;
- Identify item types (test content) that could require info/knowledge from >1 lesson/topic
- Organize time spent in class on topics – in proportion to their contribution to the overall summary of achievement that each assessment is intended to give the instructor;
- Choose assignment types
- Select/elicit test items

When developing assessments, design questions/items - within topics - keeping the target cognitive complexity in mind (from the learning objectives). This can be done with the combination of Messick's 3 questions and Bloom's taxonomy. With purposeful complexity built into the assessment, then instruction and homework can be revised to ensure that learning and assessment are both aligned (see [33]). That is, the table of test specifications can help design the assessment, but it can also help:

- Design in-class/homework questions –giving practice with the levels of complexity that are desired and intended learning outcomes of the course;
- Explicitly connect assessment/exam items with learning goals; and
- Ensure that all items on a test or assessment contribute to its quality – and the interpretability of the summary – and not just to the length of the assessment.

Using a table of test specifications, assessments can be designed in a way that groups together - purposefully - items that are exchangeable. For example, all the 1-point (factual recall) items might be considered exchangeable, so that every score of 7 out of 10 correct (for example) would be the same “type of student” as every other score of 7 – on that section of the test. Then, all the 2-point, or 3-point (etc.) items can be grouped similarly - then there would be subtotals (and if you need a single-value summary, then create a total score. Without this kind of structure, you may end up with two students with “total scores” of 70/100 that are actually *very* different: one gets all the hardest questions right but maybe didn't leave enough time/dedicate enough attention to the 1-point/memorization items (and has a total score of 70) and another student gets all the easy questions right and none of the harder questions right (and has a total score of 70). These two, qualitatively different, students cannot be distinguished by the single-value summary of the total score, because it collapses across important features of your questions.

When you organize the elements of your assessment/test using the table of test specifications and particularly, purposeful use of cognitive complexity in the items, then grouping exchangeable items together will naturally yield interpretable subscores for each complexity level. Instructors can still create a single-value summary of the overall test performance (“total score”), but identify the truly-same performances on the exam by ensuring that people with a total score of 70 did not get it by a) missing most low Bloom's level items but getting all high Bloom's items right or b) missing most high Bloom's level items but getting all lower Bloom's level items right.

These two situations can lead to the same total score, but the scorers themselves – and their learning in the course - are very different. Creating subscores of truly-exchangeable (because they are at the same level of cognitive complexity) items can make assessment scores more interpretable. Also, the *actionability* of assessment scores or grades - i.e., whether the instructor could interpret and act on whatever the assessment summary (e.g., total score) is - is important to consider, even if this is rarely done.

Quality definitions/levels of the AER: how to rate the four questions on each dimension

The AER appears in the appendix as a worksheet. With its four dimensions of any assessment being evaluated, several “rating” systems are possible. The rating approach that is selected will relate to the user of the AER rather than being a function of the AER itself. At a minimum, each feature on each dimension can be rated yes/1 or no/absent/0. A middle rating, “difficult to discern” is an important rating of “not yes/1, but also not no/missing/0”: in addition to pointing to aspects of an assessment that are not “quite finished”, when such ratings are appropriate, it can identify faculty development needs by the developer of the assessment or by the evaluator of the assessment. “revise” can mean “start over” or “this is remediable”. Thus, the AER in the appendix includes this middle rating level, although individual instructors can use or ignore it.

For simplicity, three different rating systems, reflecting different quality levels, but not very-different quality definitions, are considered:

1. Each item is rated yes/present (1) or no/absent (0). This can be extended so that, if dimension 1 is not rated a 4 (all four features present on that dimension), then none of the other dimensions are rated (i.e., all get a zero). This kind of rating system weights the first dimension most heavily.
2. Each item is rated according to how clearly the evaluator can give the rating: 1=yes/clear; .5= yes/difficult to discern <“rounding up from zero”>; 0=no/clear; and missing = difficult to discern, but probably no (can’t “round up to at least .5”).
3. Items are rated for whether or not the assessment can be/needs to be revised so that the rating is yes or yes/clear. E.g., 1=yes/clear/no revision needed; .5 = needs revision/not clear enough; and 0 = absent (and/or needs substantial revision); while a score of -1 = not revisable. This scheme would be useful if there were plans for a meeting or other opportunity to “improve assessments”, possibly also better for determination of whether mentoring/professional development is needed. Also, if any part of the faculty development plans will include the development, evaluation, and/or revision of assessments, then this scheme would help identify content/training that are needed use in those activities.

If any of these, or other, approaches to defining “levels” or scores on the AER, the decisions, values, and systematic rules for assigning the scores should be documented and shared with relevant stakeholders ([4: p. 85; Standard 4.0).

Discussion

Although they are most often utilized by developers of standardized tests, the Standards for Educational and Psychological Testing [4] describe the empirical and theoretical requirements for useful, fair, interpretable, and actionable assessments. On the first page, the Standards state: *“Well-constructed tests that are valid for their intended purposes have the potential to provide substantial benefits for test takers and test users.”* [4: p. 1]. The AER is intended to help instructors ensure that their assessments meet these validity standards. In particular, “(v)alidation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use”. [4: p. 11]. By completing the AER, an instructor can compile the evidence they have, or want to have, about the use and intended interpretation of all assessments.

The Assessment Evaluation Rubric leverages knowledge from the educational and cognitive psychological domains to help instructors ensure that their teaching is aligned with, and supportive of, their learning objectives (see [33, 34]). The clarity dimension might be prioritized last, but this is partly because the clarity of the signal any assessment provides is meaningless if alignment (dimension 1), utility (dimension 2), and cognitive complexity (dimension 3) have not already been considered. In fact, alignment of an assessment with the learning objective (dimension 1) is essential to any claim of “assessment”, so it must be prioritized first. However, the considerations of the table of test specifications suggest that ensuring that an assessment does in fact meet all of the AER dimensions is likely to be an ongoing and iterative process. The Assessment Evaluation Rubric was developed based on, and to help instructors to leverage, empirical and theoretical foundations from educational and cognitive psychology in the development and evaluation/revision of assessments in higher education. However, this foundation does not guarantee that raters using this rubric will be effective or that ratings will have high fidelity. It will be helpful if the AER is applied to a single assessment by at least two independent raters; if independent evaluators come to consensus on the ratings they give to individual assessments, those consensus-based ratings will be more believable and acceptable.

Appendix. The Assessment Evaluation Rubric

	Yes = 1 (Unambiguously, the AER dimension feature is clearly present/identifiable,)	Possibly = .5 * (Partially/possibly/seems so but can't tell; it might simply be misrepresented/mislabeled.)	No = 0 (Either there is no evidence of this dimension/feature, or whatever is available is clearly not what the AER dimension/feature requires.)
General Alignment			
1. A learning goal was articulated (that this assessment is intended to assess). ²			
2. The assessment is aligned with the learning goal(s).			
3. The teaching is supportive of what is being assessed.			
4. The assessment is aligned with the Bloom's level complexity of thinking about the content that the learner should have gained from the instruction.			
Utility of the assessment			
1. The assessment is intended to be formative.			
2. The assessment is (also) intended to be summative.			

3. The results of the assessment are informative about the learning the CURE was intended to provide.			
4. Assessment item structure is correct (e.g., common errors of multiple-choice item construction are avoided; stems do not suggest the answer; the correctness of a multi-part item is not contingent on subparts that are at diverse levels of complexity).			
Use of a formal method of incorporating complexity			
1. Bloom's (or equivalent) taxonomy is used (correctly and appropriately, i.e., not just for memorization unless that is the target cognitive level of instruction).			
2. Rationale for using, or manipulating complexity in items, is clear and coherent.			
3. There is attention given to the cognitive complexity of the items.			
4. The assessment was intended to provide the student with actionable evidence of what to do			

(better next time/to take advantage of the learning that was achieved).			
Clarity of the assessment			
1. Students would understand clearly what they're required to do "for a good grade".			
2. Instructor would be able to interpret- and act on - the summary of student work on the assessment (total score or label <pass/fail>).			
3. Items on an assessment that is summarized as a total score are exchangeable (so that a total score, if that is computed, is interpretable). Getting one item wrong provides the same information about student learning as any other item.			
4. The items on the assessment are there for a clear reason– and appropriate for the learner's level of experience, in terms of required cognitive complexity.			
NOTES: * a three-level scoring ("yes"=1, "maybe/can't tell/partially" = 0.5, and "no"=0) is shown with examples for why each of these ratings would be given. Depending on the purpose (expected results) of the AER, a simpler two level (yes/no) rating might be preferable.			

References

- 1 Walvoord BE. (2010). *Assessment Clear and Simple: A practical guide for institutions, departments, and general education*, 2E. San Francisco, CA: Jossey Bass.
- 2 Palomba CA & Banta TW. (2015). *Assessment essentials: Planning, implementing, and improving assessment in higher education*, 1E. San Francisco, CA: Jossey Bass.
- 3 Banta TW & Palomba CA. (2015). *Assessment essentials: Planning, implementing, and improving assessment in higher education*, 2E. San Francisco, CA: Jossey Bass.
- 4 American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for educational and psychological testing*, 2E. (2014). Washington, DC: American Educational Research Association.
- 5 Nilson L. (2016). *Teaching at its best: A research-based resource for college instructors*, 4E. San Francisco, CA: Jossey Bass.
- 6 Tractenberg RE, Lindvall JM, Attwood TK, Via A. (2020, April 2) Preprint. Guidelines for curriculum and course development in higher education and training. Published in the Open Archive of the Social Sciences (SocArXiv), DOI 10.31235/osf.io/7qeht
- 7 Nicholls G (2002). *Developing teaching and learning in higher education*. London, UK: Routledge.
- 8 Kirschner PA, Sweller J, Clark RE. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist* 41: 75-86.
- 9 Ambrose SA, Bridges MW, DiPietro M, Lovett MC, Norman MK. (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching*. Jossey-Bass: San Francisco, CA.
- 10 Fink LD (2013). *Creating Significant Learning Experiences: An Integrated Approach to Designing College Courses*, 2E. San Francisco, CA: Jossey Bass.
- 11 Hutchings P, Kinzie J, Kuh GD. (2015). Evidence of student learning: What counts and what matters for improvement. In GD Kuh, SO Ikenberry, & NA Jankowski (Eds.). *Using Evidence of Student Learning to Improve Higher Education*. Somerset, NJ: Jossey-Bass. Pp 27-50.
- 12 Handelsman J, Miller S, Pfund C. (2007). *Scientific teaching*. New York, NY: WH Freeman.
- 13 National Research Council. (2001). *Knowing what students know: the science and design of educational assessment*. Committee on the Foundations of Assessment. Pelligrino J, Chudowsky

N, & Glaser R Editors. Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

14 Tyler RW (1949). *Basic principles of curriculum and instruction*. Chicago: The University of Chicago Press.

15 Bloom BS (Ed), Englehart MD, Furst EJ, Hill WH & Krathwohl DR. (1956). *Taxonomy of educational objectives: the classification of educational goals, by a committee of college and university examiners. Handbook I: Cognitive Domain*. New York: David McKay.

16 Knowles MS, Holton III EF, Swanson RA. (2005). *The Adult Learner, 6E*. Burlington, MA: Elsevier

17 Messick S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23(2): 13-23.

18 Messick S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In RD Goffin, E Helmes (Eds). *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy*. Norwell, MA: Kluwer Academic Publishers. Pp. 3-20.

19 Kuh GD, Ikenberry SO, & Jankowski NA. (2016). *Using Evidence of Student Learning to Improve Higher Education*. Somerset, NJ: Jossey-Bass.

20 Weston C, Cranton PA. (1986). Selecting instructional strategies. *Journal of Higher Education*, 57(3): 259-288.

21 Lane S, Raymond MR, Haladyna TM, Downing SM. (2016). Test development process. In S Lane, Raymond MR, & Haladyna TM (Eds). *Handbook of Test Development, 2E*. New York, NY: Routledge. Pp. 3-18.

22 National Institute for Learning Outcomes Assessment (NILOA, 2016, May). Creating sustainable assessment through collaboration: A national program reveals effective practices. http://learningoutcomesassessment.org/documents/Occasional_Paper31.pdf

23 Dawson P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education* 42(3): 347-360.
<https://doi.org/10.1080/02602938.2015.1111294>

24 Stevens DD & Levi A. (2005). *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning, 2E*. Sterling, VA: Stylus Publishing.

25 Wiggins G. & McTighe J. (2005). *Understanding by Design, 2E*. New York, NY: Pearson.

- 26 Moseley D, Baumfield V, Elliott J, Gregson M, Higgins S, Miller J, Newton DP. (2005). *Frameworks for Thinking: A handbook for teaching and learning*. Cambridge, UK: Cambridge University Press.
- 27 Tractenberg RE, Gushta MM, Mulroney SE, Weissinger PA. (2013) Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Advances in Health Sciences Education* 19: 945-961.
- 28 Jensen JL, McDaniel MA, Woodard, SM, & Kummer TA (2014). Teaching to the test...or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, 26(2), 307-329. doi:10.1007/s10648-013-9248-9.
- 29 Tractenberg RE. (2017). How the Mastery Rubric for Statistical Literacy can generate actionable evidence about statistical and quantitative learning outcomes. *Education Sciences. Special Issue: Consequential Assessment of Student Learning*. 7(1), 3; doi:10.3390/educsci7010003 <http://www.mdpi.com/2227-7102/7/1/3/pdf>
- 30 Fives H and DiDonato-Barnes N. (2013). Classroom Test Construction: The Power of a Table of Specifications. *Practical Assessment, Research, and Evaluation* 18(3) Article 3: 7 pages. DOI: <https://doi.org/10.7275/cztt-7109> downloaded from <https://scholarworks.umass.edu/pare/vol18/iss1/3> on 14 April 2020.
- 31 Frey BB. (2018). "The Table of Test Specifications". In, *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Thousand Oaks, CA: Sage. Article accessed at <https://dx.doi.org/10.4135/9781506326139.n685> on 14 April 2020.
- 32 Tractenberg RE. (2019, October 11). The Table of Test Specifications: mapping test domains and content with cognitive complexity and learning goals. Available (annotated) at: https://www.academia.edu/40610442/The_Table_of_Test_Specifications_ToTS
- 33 Hutchings, P. (2016). *Aligning educational outcomes and practices*. (Occasional Paper No. 26). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.
- 34 William D & Thompson M. (2008). Integrating Assessment with Instruction: What will it take to make it work? In C. A. Dwyer (ed.). *The Future of Assessment: Shaping Teaching and Learning*. London, UK: Routledge. pp. 53–82.