education sciences

MDPI

Article

The Assessment Evaluation Rubric: Promoting Learning and Learner-Centered Teaching through Assessment in Face-to-Face or Distanced Higher Education

Rochelle E. Tractenberg

Collaborative for Research on Outcomes and Metrics: Departments of Neurology, Biostatistics, Biomathematics, and Bioinformatics, and Rehabilitation Medicine, Georgetown University, Washington, DC 20057, USA; rochelle.tractenberg@gmail.com

Abstract: It is common to create courses for the higher education context that accomplish content-driven teaching goals and then develop assessments (quizzes and 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 describes, or permits the evaluation of, 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, and clarity) has four questions that can be rated as present/absent. Other rating methods can also be conceptualized for 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. As instructors shift from face-to-face toward virtual or hybrid teaching models, or as they shift online instruction (back) to face-to-face teaching, it creates an ideal opportunity to ensure that assessment is optimizing learning and is valid for instructional decision-making.

Keywords: higher education; assessment; instructional development; validity

Citation: Tractenberg, R.E. The Assessment Evaluation Rubric: Promoting Learning and Learner-Centered Teaching through Assessment in Face-to-Face or Distanced Higher Education. *Educ. Sci.* **2021**, *11*, 441. https://doi.org/10.3390/educsci11080441

Academic Editors: Kim Koh and Olive Chapman

Received: 4 May 2021
Accepted: 4 August 2021
Published: 18 August 2021

1. Introduction

Assessment is formally 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" (p. 2, [1]; see also p. 4, [2]; pp. 1–13, [3]). Walvoord [1] clarifies 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; p. 75, [4]). These sources agree that assessment should be 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 can and should be intimately tied to the objectives of the instruction (p. 15, [3,5]). Tractenberg et al. (2020) [6] discuss a five-phase model for curriculum and instructional development (based on [7]) explaining how the alignment of objectives and assessment can strengthen the likelihood of the desired learning. In this model [6,7], the "assessment" phase (4 of 5) explicitly considers the relationship of assessment to the learning objectives defined in phase 1; how teaching is accomplished (phase 2) is also tied to the assessment. As instructors in higher education consider face-to-face, virtual, or

hybrid teaching models, or as they shift their courses from one to the other model, an ideal opportunity arises to ensure that assessment is optimizing learning.

Assessment is defined as *formative* when decisions based on it 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 for the given cohort (p. 19, [3]). From the perspective of students in higher education, 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; it is clearly how they themselves are rated or ranked at the end of a course. However, decisions that affect student learning, if based on summative assessments, are only able to affect the learning of *later* students, whereas formative assessments could be used to affect the learning of the *current* group of students. When assessment is effectively integrated within and during a course or curriculum, it can be used to make decisions about how best to teach in order to optimize student learning. This is true irrespective of the mode of the instruction; in the post-COVID-19 era, ensuring that assessment is aligned with instructional (learning) objectives can ease transitions to and from different modalities within a school term, year, or program. 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). Moreover, Kirshner et al. (2006) [8] note, "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].

Formative assessment represents at least part of this "guidance of the student learning process", although summative assessment tends to be more widely used. This is because a "whole class" or course segment can be easily described with the distribution of test scores, whereas formative assessment will be idiosyncratic and challenging to summarize over students and over time. 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. The adult education, educational psychology, and cognitive science domains have long documented what works to promote learning [9]. A key feature of a successful curriculum in higher education is that it contains opportunities for assessment that both enable learners 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 (p. 28, [13]).

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 are central to all 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.

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; p. 20, [13]). Handelsman et al. (2007) [12] point out that most higher education faculty do not have much or 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 the relevant literature, converging on several straightforward principles. The observation by Handelsman et al. [12] suggests that what is *common practice* in assessment across much of higher education may not be adequately aligned with what is "best practice" according to the learning sciences. The purpose of this paper is to describe a new tool, the Assessment Evaluation Rubric (AER), that synthesizes these best practices in a single source that can be used across higher education contexts (time in program and disciplines and assessment types). This tool was developed to make the relevant disciplinary knowledge from the learning sciences more easily accessible for both the creation and the evaluation of assessments in higher education. As higher education instructors contemplate their teaching moving forward in the aftermath of changes forced by COVID-19, the AER can be a useful tool to apply to assessments, whether existing or to be developed. Instructors thinking about moving from online back to in-class instruction can take the opportunity to ensure their assessments are aligned with their learning objectives and that the learning objectives are concrete and observable. The AER encourages those creating or evaluating their assessments to consider four key characteristics or dimensions:

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

All assessments can be described along these four dimensions, which originate from current authorities on assessment in education [3–5,10,17–21]. 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. Validity is defined (psychometrically) as "the extent to which evidence and theory support the use/interpretation of the summary of the assessment" [4]. A lack of interpretability compromises the validity of the assessment, meaning it cannot consistently support the instructors' decisions. The AER was developed to capture these four dimensions in a concise way, to guide both the development and evaluation/revision of assessments using these four dimensions in higher education contexts. The AER, and this paper, serve to synthesize the multitude of authoritative references that most university instructors would never have occasion to read. This new four-dimensional rubric can be used by the individual for themselves or by teams/for others, is novel, and is formulated according to longstanding findings about learning and assessment from the cognitive and educational sciences.

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) *while* that teaching and learning are happening [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" (p. 6, [22]). Effective formative assessment is, ideally, immediately 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 as they are happening 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 to either teaching or learning of the knowledge, skills, and abilities that were summatively assessed. Summative assessment may be used to document student achievement, and as such, is less actionable than formative assessment is or can be. By describing students after the end of an instructional opportunity, summative assessment may be less actionable for institutions than formative assessment is.

As instructors consider shifting instruction—and assessment—from in-person to online and/or to hybrid situations, or back, it offers an excellent opportunity to consider the arguments and decisions they want to make about their students' learning.

2. 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. The four dimensions, listed above, pertain in all educational contexts, but this rubric and the discussion below were developed specifically for higher education.

Using Dawson's (2015) [23] rubric design element vocabulary, five of the 14 rubric design elements that are most relevant for the AER are task-specificity, quality definitions, secrecy, scoring strategies, and quality levels.

The Assessment Evaluation 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 object of the AER is thus *task-specific* where the task is the creation or evaluation 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). Conversely, 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, and the result of applying it to an assessment, is meant to be shared widely, as it represents a synthesis of empirical work on education and the validity of actionable decisions in higher education. Thus, the AER is intended to make the research findings about assessment from the learning sciences easily available to every instructor.

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. The AER is essentially a framework for a Degrees of Freedom Analysis [25], specifically intended to support educational decision-making [26] around the design and use of a specific assessment. Because the AER 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.

Each of the four dimensions has four questions, which were culled from either the source authorities or formulated for clarity and concision. Each dimension, with its four aspects, is listed and discussed below.

The first of the four dimensions to the AER is general alignment. This dimension is intended to capture the relevance of the assessment for the instructor (see, e.g., [3,10,17]).

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 (or similar) level complexity of thinking about the content that the learner should have gained from the instruction.

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 are tied to Messick's three features of valid assessment [17]:

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

These questions are also embedded in Wiggins and McTighe's Backwards Design (2005) [27] and appear in other well-established cognitive psychological principles. Whether instruction is face-to-face, distanced, or a hybrid, these principles pertain.

Whatever tasks students are asked to carry out in the assessment (i.e., Messick #3) should have some relationship with *observable behaviors*—because to accomplish Messick #2, the actions that students perform 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 [28]). Keeping the three Messick questions in mind as an assessment is developed or revised will make it more likely to achieve/satisfy all of the features in the alignment dimension. These three Messick questions may also be essential for shifting from face to face, to online, to hybrid, or between these instructional modalities. In changing instructional modalities, it is unlikely that the KSAs have changed, but the actions that reveal the accomplishment of the target learning, and the tasks to elicit those behaviors, may very well have to change with instructional modality. While not all cognitive processes are observable, only those that *are* observable can be assessed reliably at scale. Bloom's Taxonomy (or other cognitive taxonomies) allows the identification of specific observable behaviors and applying Messick's criteria ensures that assessment will result in those behaviors, and they will be *observable*, through the structure of the assessment.

The second dimension of the AER is utility. As noted, a valid assessment is one with strong evidence that supports the interpretation of its scores (or summary) ([4]; see also [3,18,19]). Thus, understanding how the assessment is intended to be used is an essential feature.

Dimension 2. Utility/use 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, and the correctness of a multi-part item is not contingent on subparts that are at diverse levels of complexity).

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 AER 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 receives 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 the utility 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 the alignment dimension 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. Good multiple-choice items are notoriously difficult to write, particularly at a specific cognitive complexity level [29]. 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 ratings on the alignment dimension are low (or all 0), the utility of the assessment will similarly be low or lacking. If alignment dimension ratings are low and somehow ratings on the utility dimension are higher, this will point to misinterpretations of the AER dimensions rather than an assessment that is *not* aligned with instructional purposes/objectives but is still a useful assessment.

Once alignment and utility features are fully and consistently present for an assessment, then a deeper examination of the role of cognitive complexity (Chapter 1, [5]; p. 213, [10,20]) in the structure of the assessment can fruitfully be examined (Dimension 3).

Dimension 3. Alignment of the assessment's Bloom's (or similar) level cognitive 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 of 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 feature 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 [29,30]). If the instruction is geared toward 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 [29]). 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 the cognitive complexity dimension can promote an emphasis on learn*ing*, and the learn*er*, rather than on teaching and the teacher, throughout instruction. However, the suitability of the cognitive complexity of assessment items is not a meaningful determination if the assessment is not aligned with learning goals (Dimension 1) or useful for making decisions about learning (Dimension 2).

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., [31]). 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. To change or confirm the cognitive complexity of an assessment, instructors can utilize Bloom's taxonomy ([15], see also [29]) or any relevant framework from the compendium compiled by Moseley et al. (2005) [28]. 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 Chapter 1, [5]; p. 213, [10,20]).

There are many cognitive complexity taxonomies (see [28]). In the United States, Bloom's [15] is the most common/familiar one. A selected taxonomy should be utilized for both developing an assessment (or constituent items) and the learning outcomes the assessment supports. Based on experience, or simplicity of scoring decisions, instructors may believe they "need to have 10 items on a quiz", but if the learning objective is to "encourage critical thinking", asking students to demonstrate their memorization of ten facts *cannot support that objective*. Critical thinking cannot be assessed easily with multiple-choice exams, because this item type is most consistent with memorization and recognition. Critical thinking requires higher cognitive sophistication, but essays and even short-answer assessment items can be hard to grade consistently across responses. However, these are not coherent reasons for why ten multiple-choice items would be included in any assessment where critical thinking is to be taught or practiced. As instructors contemplate the similarities and differences between in-person, online, and hybrid instruction, they can also consider whether the assessment method they are using is better suited to one or the other context. The AER dimensions can help instructors think through these considerations.

Finally, items on an assessment may vary widely in the cognitive complexity they require. The rationale for that variability of complexity—even if systematic—might not be appreciable to the learner. This could undermine the perceived coherence of the assessment. In fact, coherence in the content of the assessment is the focus of the fourth AER dimension, clarity [4,21].

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]). The clarity dimension appears last because, without the other dimensions, perfect "clarity" does not make the assessment useful. Modifications to improve clarity can include reordering items, integrating examples or better/clearer instructions, and other, similar information that informs 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 assessment for the learner. These questions can also strengthen the instructor's ability to interpret the "score" or performance on any assessment.

3. Assessment Construction or Revision Using Tables of Test Specifications

As an assessment is constructed, or evaluated, using the AER, instructors may also take the opportunity to map out exactly what information they hope to obtain from each test. One approach to ensuring that a test will generate an interpretable signal to the instructor about the learner's achievement is a table of test specifications or test blueprint [32–34]. A table of test specifications is a table that describes the features of your test. It lists the contents, distribution of questions (number, number per topic), cognitive complexity, and/or item type. As such, a table of test specifications can be extremely informative for instructors about their assessment and the decisions it can help them make about their students. Such a table can be constructed for any type of assessment; this structure (mapping to course topics or learning objectives) can also be applied to an existing rubric (for qualitative assessment). Table 1 shows a simple representation of the coverage of five topics on a single test.

Table 1. Table of test specifications showing distribution of questions or points on a hypothetical 100-item exam covering five topics.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Whole Test
30%	12%	40%	9%	9%	100%

A table like Table 1 may surprise instructors by the imbalance of items about one topic, or it might show exactly the distribution that these topics need for the intended use of the test scores. Ensuring that students understand that the test will be more heavily weighted toward Topic 3 would improve the clarity of the assessment (and would help students to plan study time). If Table 1 represents scoring, rather than item count, there could be only one question on Topic 1, which represents 30% of the total score or grade (knowledge that would also help students to study).

Table 2 shows the distribution of item *types* within a single test, which corresponds more to the instructor's attention to Dimension 3 (cognitive complexity) than to content or topics, which is the sole focus in Table 1. Table 2 would also help an instructor ensure that specific learning objectives (Dimension 1) were met, especially if the learning objectives involve cognitive sophistication. That is, 30% of the grade will require higher-order cognitive processing (i.e., for essays), while 58–70% of the grade will reflect the lowest Bloom's level of complexity (recognition and memorization). The short-answer questions may be structured to require application, prediction, some analysis, or illustration, putting them at higher cognitive levels than recognition and memorization.

Table 2. Table of test specifications showing cognitive complexity demands on a hypothetical 100-point assessment.

Essay	Short Answer	Multiple Choice Question (MCQ)	Fill In	Matching	Whole Test
30%	12%	40%	9%	9%	100%

Item type and topic can then be combined into a single table that describes the specifications of one test (where row totals are 100% of items on that topic). Table 3 shows that the test requires high cognitive complexity in the form of essay assessment items for each topic, and that recognition and memorization (low cognitive complexity) are contributing the majority of information to the instructor's decisions about each student.

Table 3. Table of test specifications with topical and cognitive complexity coverage.

	Essay	Short Answer	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 aligned with the learning objectives for the course and also that they include item types that require the targeted level of cognitive complexity, as shown in Table 4.

Table 4. Table of test specifications with learning objectives and item types.

Essay	Short Answer	MCQ	Fill In	Matching
30%	12%	40%	9%	9%
Objectives 1, 4	Objective 2	Objectives 1–4	Objective 3	Objective 4

Table 4 can be constructed for each topic to ensure that learning objectives (if articulated separately by topic in one course) are aligned with assessment (AER Dimension 1). The instructor's description of item type and learning objectives can strengthen decision-making about the test scores (AER Dimension 2). Knowing that the essay requires the highest cognitive sophistication, followed by short-answer, with the other item types all at the lowest cognitive complexity levels, will help instructors understand how they can use the test scores to make decisions about student achievement (if used summatively) or about possible changes to make to teaching in order to help students achieve more in the learning that follows this assessment (if used formatively).

In keeping with AER Dimension 3, instructors may consider how cognitively complex (in Bloom's, or B, levels) the assessment of learning on each topic will be in the assessment.

Table 5 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. Since there is only 1% of items (i.e., one of 100 items) at B4, and this item assesses Topic 1 only, this table shows that there might be a mismatch between the instructor's learning goals and how they are being assessed. Table 5 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 [35,36]. When developing assessments, instructors should design questions/items—within topics—keeping the target cognitive complexity in mind. With purposeful complexity built into the assessment, then instruction and homework can be revised to ensure that learning and assessment are both aligned (see [35]).

Table 5. Table of test specifications with Bloom's (B) level representation by topic.

	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%

4. Quality Definitions/Levels of the AER: How to Rate the Four Questions on Each Dimension

Because the AER is a true rubric, there is a flexibility in exactly how, and with what content, an assessment can exhibit the four dimensions. 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; such frameworks can be combined with/utilized as the learning objectives in AER Dimension 1, the identification of and alignment of the assessment with specific learning objectives about which the assessment is intended to provide evidence.

The AER appears in Appendix A as a worksheet. Each question can be rated yes/1 or no/absent/0. A middle rating could be included, e.g., "difficult to discern" or "more detail needed": This middle rating can also identify faculty development needs among the assessment developers, needing training to strengthen assessment construction, or among evaluators, needing training to strengthen understanding of the four AER dimensions. Thus, the AER in Appendix A includes this middle rating level, and individual instructors can use, revise, or ignore it. Examples of rating options include:

1. Each item is rated yes/present (1) or no/absent (0). This can be extended so that, if the alignment dimension is not rated a 4 (all four features present on that dimension), then none of the other dimensions are rated (i.e., all receive 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). A rating of −1 = not revisable/start over. This scheme would be useful if there were plans for a meeting or other opportunity to "improve assessments" and possibly also better for the determination of whether mentoring/professional development is needed. Additionally, 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 and training that are needed in those activities.

If any of these, or other, approaches to defining "levels" or ratings on the AER, the decisions, values, and systematic rules for assigning the scores should be documented and shared with relevant stakeholders (p. 85, [4]; Standard 4.0). Moreover, self-assessment using the AER can be performed using qualitative ratings such as "sufficient" and "get help to revise/improve", the quest for ensuring an assessment has "sufficient" ratings on all four dimensions can easily be discussed in a teaching portfolio or as a professional development goal for instructors at any level. Mentor instructors can also use the AER (with any rating scale) to facilitate both learning-centered assessment and the professional portfolio development that instructors may need for advancement.

5. Discussion

The AER was created based on what is known about learning and learners in higher/adult education: Bloom's Taxonomy of Cognitive Behaviors [15], which allows the identification of 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 the desired observable and interpretable behaviors by students; and principles of andragogy [16], to ensure that adult learners are "getting" what the curriculum seeks to deliver. The four AER dimensions can either be applied to an existing assessment, or they can be used to support the construction of new assessments to provide valid information about both learning and instruction. As instructors consider shifting from or to online/distance or hybrid instructional modalities involving some face-to-face and some distance learning, they may also be interested in ensuring that their assessments are consistent with their learning objectives. The AER can be used for any of these purposes.

The AER is an instantiation of a semi-qualitative method of analysis, Degrees of Freedom, originally published in 1975 [25] but modified in 2019 to specifically support educational decision-making [26]. It is intended to bring the authoritative work in learning sciences together into a single resource for considering, creating, or revising assessments. Whether summative or formative, assessment is meant to inform decision-making, and as such, the Standards for Educational and Psychological Testing [4] should be leveraged, if not utilized explicitly. "*Assessment* is a broader term than *test*, commonly referring to a process that integrates test information with information from other sources." (p. 2, [4] emphasis in original). 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" (p. 11, [4]).

Thus, the AER can be used by instructors to describe, and better understand their decisions about, student learning. It can also provide a concrete method to link (or ensure a link between) instruction and assessment, particularly with Dimension 1. Additionally, because it has general questions that diverse assessments can be rated on, the AER can be used to test or study if multiple assessments are similar—or different—in educational research. The AER can be used to examine assessments from disparate instructors and domains (and for learners at different levels). The AER can be used for actionable evaluation of both formative and summative assessments.

The AER 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 [35,36]). The clarity dimension might be prioritized last in the AER, 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. Ensuring that an assessment does in fact meet all of the AER dimensions is likely to be an ongoing and iterative process. It will be helpful if the AER is applied to a single assessment by at least two independent raters; if independent evaluators come to a consensus on the ratings they give to individual assessments, those consensus-based ratings will be more believable and acceptable.

Importantly, it may not be feasible to expect that all evaluators (users of the AER) will know what the "correct" answer to AER items is—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.

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"

(p. 1, [4]). 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" (p. 11, [4]). 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.

Funding: 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). APC was waived for this article.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. The Assessment Evaluation Rubric Worksheet.

	Yes/Present = 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. This Rating Option May Be Omitted)**	**No/Absent = 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. 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 cognitive complexity 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 instruction 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).			

Table A1. *Cont.*

	Yes/Present = 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. This Rating Option May Be Omitted)	No/Absent = 0 (Either There Is No Evidence of This Dimension/Feature, or Whatever Is Available Is Clearly Not What the AER Dimension/Feature Requires)
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. The 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 are 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: * Depending on the user's purpose using the AER, a simpler two level (yes/no) rating might be preferable.

References

1. Walvoord, B.E. *Assessment Clear and Simple: A Practical Guide for Institutions, Departments, and General Education*, 2nd ed.; Jossey Bass: San Francisco, CA, USA, 2010.
2. Palomba, C.A.; Banta, T.W. *Assessment Essentials: Planning, Implementing, and Improving Assessment in Higher Education*, 1st ed.; Jossey Bass: San Francisco, CA, USA, 2015.
3. Banta, T.W.; Palomba, C.A. *Assessment Essentials: Planning, Implementing, and Improving Assessment in Higher Education*, 2nd ed.; Jossey Bass: San Francisco, CA, USA, 2015.
4. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*, 2nd ed.; American Educational Research Association: Washington, DC, USA, 2014.
5. Nilson, L. *Teaching at Its Best: A Research-Based Resource for College Instructors*, 4th ed.; Jossey Bass: San Francisco, CA, USA, 2016.

6. Tractenberg, R.E.; Lindvall, J.M.; Attwood, T.K.; Via, A. Guidelines for curriculum and course development in higher education and training. *Open Arch. Soc. Sci.* **2020**. [CrossRef]
7. Nicholls, G. *Developing Teaching and Learning in Higher Education*; Routledge: London, UK, 2002.
8. Kirschner, P.A.; Sweller, J.; Clark, R.E. Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educ. Psychol.* **2006**, *41*, 75–86. [CrossRef]
9. Ambrose, S.A.; Bridges, M.W.; DiPietro, M.; Lovett, M.C.; Norman, M.K. *How Learning Works: Seven Research-Based Principles for Smart Teaching*; Jossey-Bass: San Francisco, CA, USA, 2010.
10. Fink, L.D. *Creating Significant Learning Experiences: An Integrated Approach to Designing College Courses*, 2nd ed.; Jossey-Bass: San Francisco, CA, USA, 2013.
11. Hutchings, P.; Kinzie, J.; Kuh, G.D. Evidence of student learning: What counts and what matters for improvement. In *Using Evidence of Student Learning to Improve Higher Education*; Kuh, G.D., Ikenberry, S.O., Jankowski, N.A., Eds.; Jossey-Bass: Somerset, NJ, USA, 2015; pp. 27–50.
12. Handelsman, J.; Miller, S.; Pfund, C. *Scientific Teaching*; WH Freeman: New York, NY, USA, 2017.
13. National Research Council. *Knowing What Students Know: The Science and Design of Educational Assessment*; National Academy Press: Washington, DC, USA, 2001.
14. Tyler, R.W. *Basic Principles of Curriculum and Instruction*; The University of Chicago Press: Chicago, IL, USA, 1949.
15. Bloom, B.S. *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*; David McKay Co Inc.: New York, NY, USA, 1956.
16. Knowles, M.S.; Holton, E.F., III; Swanson, R.A. *The Adult Learner*, 6th ed.; Elsevier: Burlington, MA, USA, 2005.
17. Messick, S. The interplay of evidence and consequences in the validation of performance assessments. *Educ. Res.* **1994**, *23*, 13–23. [CrossRef]
18. Messick, S. Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy*; Goffin, R.D., Helmes, E., Eds.; Kluwer Academic Publishers: Norwell, MA, USA, 2000; pp. 3–20.
19. Kuh, G.D.; Ikenberry, S.O.; Jankowski, N.A. *Using Evidence of Student Learning to Improve Higher Education*; Jossey-Bass: Somerset, NJ, USA, 2016.
20. Weston, C.; Cranton, P.A. Selecting instructional strategies. *J. High. Educ.* **1986**, *57*, 259–288. [CrossRef]
21. Lane, S.; Raymond, M.R.; Haladyna, T.M.; Downing, S.M. Test development process. In *Handbook of Test Development*, 2nd ed.; Lane, S., Raymond, M.R., Haladyna, T.M., Eds.; Routledge: New York, NY, USA, 2016; pp. 3–18.
22. National Institute for Learning Outcomes Assessment (NILOA). (2016, May); *Higher Education Quality: Why Documenting Learning Matters*; University of Illinois and Indiana University: Urbana, IL, USA, 2016. Available online: https://files.eric.ed.gov/fulltext/ED567116.pdf (accessed on 9 June 2016).
23. Dawson, P. Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assess. Eval. High. Educ.* **2017**, *42*, 347–360. [CrossRef]
24. Stevens, D.D.; Levi, A. *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*, 2nd ed.; Stylus Publishing: Sterling, VA, USA, 2005.
25. Campbell, D.T. "Degrees of Freedom" and the Case Study. *Comp. Political Stud.* **1975**, *8*, 178–193. [CrossRef]
26. Tractenberg, R.E. Degrees of Freedom Analysis in educational research and decision-making: Leveraging qualitative data to promote excellence in bioinformatics training and education. *Brief. Bioinform.* **2019**, *20*, 416–425. [CrossRef] [PubMed]
27. Wiggins, G.; McTighe, J. *Understanding by Design*, 2nd ed.; Pearson: New York, NY, USA, 2005.
28. Moseley, D.; Baumfield, V.; Elliott, J.; Gregson, M.; Higgins, S.; Miller, J.; Newton, D.P. *Frameworks for Thinking: A Handbook for Teaching and Learning*; Cambridge University: Cambridge, UK, 2005.
29. Tractenberg, R.E.; Gushta, M.M.; Mulroney, S.E.; Weissinger, P.A. Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Adv. Health Sci. Educ.* **2013**, *19*, 945–961. [CrossRef] [PubMed]
30. Jensen, J.L.; McDaniel, M.A.; Woodard, S.M.; Kummer, T.A. Teaching to the test...or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educ. Psychol. Rev.* **2014**, *26*, 307–329. [CrossRef]
31. Tractenberg, R.E. How the Mastery Rubric for Statistical Literacy can generate actionable evidence about statistical and quantitative learning outcomes. *Educ. Sciences. Spec. Issue: Consequential Assess. Stud. Learning* **2017**, *7*, 3. [CrossRef]
32. Fives, H.; DiDonato-Barnes, N. Classroom Test Construction: The Power of a Table of Specifications. *Pract. Assess. Res. Eval.* **2013**, *18*, 7. Available online: https://scholarworks.umass.edu/pare/vol18/iss1/3 (accessed on 14 April 2020). [CrossRef]
33. Frey, B.B. The Table of Test Specifications. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*; Sage: Thousand Oaks, CA, USA, 2018; Available online: https://dx.doi.org/10.4135/9781506326139.n685 (accessed on 14 April 2020).
34. Tractenberg, R.E. The Table of Test Specifications: Mapping Test Domains and Content with Cognitive Complexity and Learning Goals. Available online: https://www.academia.edu/40610442/The_Table_of_Test_Specifications_ToTS_ (accessed on 11 October 2019).
35. Hutchings, P. *Aligning Educational Outcomes and Practices. (Occasional Paper No. 26)*; University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment: Urbana, IL, USA, 2016.
36. Wiliam, D.; Thompson, M. Integrating Assessment with Instruction: What will it take to make it work. In *The Future of Assessment: Shaping Teaching and Learning*; Dwyer, C.A., Ed.; Routledge: London, UK, 2008; pp. 53–82.