JOURNAL OF APPLIED MEASUREMENT, *24*(1/2), 1-8

FOCUS ARTICLE

Clarifying Inputs and Outputs of Cognitive Assessments

William D. Schafer
University of Maryland

This paper is about actions testing programs can take to improve communications about cognitive tests with interested stakeholders, including what they measure and the uses of the results. Concern about testing in the United States has often centered around focusing too much on rote learning and narrowing the curriculum to only tested concepts and procedures, though more recently, there is concern as well about perceived bias in outcomes. I agree that acceptance of cognitive assessments will be more likely if we can improve public understanding of them and their usefulness in ways that go beyond the limited (although important) concerns about emphasis and fairness. But what can the assessment industry do to enhance transparency? I suggest here that (1) greater attention to and clarification about the domains of major standardized tests in relation to curriculum and (2) access to interactive devices to interpret their outcomes would help the public focus on what tests actually are and what they can do. I suggest improved understandings of what tests assess, how they assess, and what the results imply for both individuals and groups can move public conversation toward ways of meaningfully addressing (and studying) assessment concerns. I suggest (1) an approach to clarifying the domain of an assessment, useful for both test developers and examinees as well as other audiences, such as instructors; and (2) a web-based means for users to tailor contextualization of results for persons and for groups using both norm- and criterion-referenced information. Although these two concepts are presented only to convey feasibility, I suggest that using processes like them will foster better-focused tests and enable more effective use of the results.

Keywords: heuristics, domain description, test interpretation

Requests for reprints should be sent to William D. Schafer, 12523 Camino Vuelo, San Diego, CA 92128, USA; email: wschafer@umd.edu

Introduction

This paper describes approaches that seem to me to hold promise for improving our ability to capitalize on two fundamental purposes of cognitive assessments: referencing a well-defined group of achievements and generating useful insights from performance. These suggestions stem from my background as a statewide director of student assessment, a faculty member at a major university, and a frequent consultant on statewide assessment following my retirement. First, I will describe a process for defining an assessed domain that could be useful for test developers, for educational professionals designing preparation programs, and for examinees themselves as they focus their own activities. Second, I will discuss an approach that might be used to facilitate both norm-referenced and criterion-referenced insights into observed test performance for individuals and for groups. The context for my suggestions is large-scale standardized cognitive assessments, although extensions to other contexts may be considered in the future. I will cite sources where I have further elaborated on aspects of these suggestions as available.

In the United States (US), societal distrust of assessments seems most to center around standardized admissions testing, although other cognitive assessments, such as secondary and even elementary school programs, particularly with high stakes, have been criticized. Some of these criticisms have stemmed from standardization, such as narrowing of the curriculum and emphasizing rote learning over higher-order thinking. Other criticisms have centered around issues of perceived bias. For example, Koljatic et al. (2021) have recently identified a clear problem of bias in public understanding of assessment outcomes, whether it exists in reality or in perception (possibly both). Although their article focused on college admissions testing, their analyses apply quite well to other standardized tests, especially accountability assessments in schools and admissions or licensing testing for advanced education and career programs. Their article was the focus of a special issue of *Educational Measurement: Issues and Practice*, with commentary from several professionals. Of these, only one seems to suggest a direction for fundamental change in what assessment professionals actually do, and that change was to make the full item pool available prior to the test, which is presumably a sample from that pool (Albano, 2021).

An example of a perhaps more feasible and, in several ways, more useful approach to the problem of test description is given here. Its basis is an extension of the well-known concept of behavioral objectives.

The second suggestion in this paper centers around test interpretation. Test results are too often communicated in ways that raise more questions than they answer. Virtually all these questions have to do with contextualization. What does a score mean about what an examinee can do? How does a score compare with scores of various groups of examinees? Is the score close to a boundary that implies sufficiency of outcome (e.g., a proficiency level or a cut-off point or a grade)? Is performance in one sub-area markedly different than another? How do my students compare with another instructor's students? Do my employees' outcomes differ markedly from each other? Which schools are more exemplary, and which need more help? Are scores changing over time? These are examples of legitimate questions that test data can inform and eventually result in improved outcomes throughout any organization using the assessment program(s). However, each requires additional information beyond the test score (i.e., a context), and addressing them can often prove impossible in practice or, when possible, is, in many cases, cumbersome at best and misleading at worst. A user-friendly, web-based process for score interpretation for persons and groups is described here.

The two proposals in this paper are presented not as recommendations but as "thought examples" that demonstrate feasibility. Alternate methodologies nevertheless could be compared with these processes to evaluate whether they enable: (1) clear and agreed-upon domain descriptions meeting the needs of multiple users; (2) clarified processes for the construction of tests that represent their domains; (3) contextualization of results for individuals and groups using both criterion- and norm- referencing; and (4) the ability of users to tailor contexts to investigate reasonable questions they may have using the data.

Specifying the Test Domain

We should not be surprised that the public is suspicious of assessments. Anyone distrusts what he or she does not understand. Even the release of the test does not convey information about how (or how well) the test represents its domain and how well its domain represents the curriculum, and as a by-product so-called "item-bashing" becomes a typical (and some might say amusing) but fundamentally useless activity.

Much like the language of sampling of people, we can think of three levels of content sampling. Like the intended population, the curriculum is the full body of knowledge and skills that might (ideally, should) be covered in a unit of instruction. As the available population, the domain of the test is a subset of the curriculum (though it might be the full curriculum, of course). Finally, like the actual sample, the test itself is the operational representation of the domain. The connections among these three, curriculum, domain, and test, should be clear, and how they are aligned should be specified and communicated as appropriate.

The domain of an assessment perhaps can be the most helpful place to begin to create an understanding about what a test or testing program is intended to measure. If we do not communicate the domain, in effect, we are saying, "We are going to test you (perhaps with a very high-stakes test), but we are not going to tell you what the test will cover." That does nobody any good, even the assessment community, since we are also testing the ability of the examinees (and/or their instructors) to guess about the nature of the domain. We need a way to communicate the domain to all relevant public groups as well as examinees without compromising test security. With that as the goal, I describe in this section an approach to domain description that should satisfy any reasonable audience. I have tried to extend the usual concept of a behavioral objective to enhance communication of the breadth, depth, and limitations of each section of the domain. The approach is presented not as a finished-product recommendation but as a demonstration that improvements in understanding assessment inputs on the part of the public are possible, and I feel it would be a useful device for test developers and instructors, as well.

Any task an examinee is asked to perform on a test has at least two parts: the content (what the examinee is asked to know and/or use) and the activity (what the examinee is asked to do with the content). There are several taxonomies for the latter in order to elaborate what these are for an assessment as well as for instruction. The revision of the familiar Bloom taxonomy by Anderson and Krathwohl (2001) is a useful example for two reasons: it is easier to understand and use, and it gives a taxonomic structure for the knowledge dimension as well as the cognition dimension. There are several others. Indeed, it is not clear even that all content areas should use the same taxonomy. But what does seem clear is that education should be toward using and not merely knowing concepts and relationships. Thus, the activity (thinking, cognition) portion of a task deserves to be elevated in importance; the content domain of an assessment should not be (but too often is) merely a list of topics.

These two parts of a behavioral objective (Anderson & Krathwohl, 2001), content and activity, can be thought of as fundamental to any achievement, but for purposes of communicating curricula, as intended for instruction or for testing, further elaboration is needed. It is suggested here that the content and activity combinations be generalized and augmented to define what has been called

"heuristics" by Schafer and Moody (2004). Heuristics are intended to be general enough to encompass many assessment opportunities (e.g., test items) but specific enough to (a) make obvious whether any given assessment prompt is within its scope as well as (b) limit the content and processing with which the heuristic will be assessed (called "assessment limits," a term coined by the Maryland high school assessment team). For example, a heuristic in an algebra assessment might be to solve a system of linear equations with up to two unknowns, where solving is the activity, systems of equations are the content, and algebraic manipulation and linearity, as well as no more than two unknowns, are the assessment limits. Another example could be to transform a word problem involving rate, time, and distance into a system of one or two linear equations amenable to a solution. Clearly, while there are an unlimited number of assessment (and instructional) opportunities for either of these, the suitability of any of them can be determined clearly for either heuristic as within or outside its scope (if the task fails, the heuristic can provide directions for revision).

Please note in passing the usefulness of statements like this for curriculum and instruction, for studying, and for providing direction to test and item authors. Taken together, they can become an "at-least" list for instructional and other preparation activities and an "at-most" list for test developers. Additionally, they can be circulated widely and debated by stakeholders to arrive at a mutually agreed-upon curriculum taken together and their individual suitability as having the right (useful) level of specificity and generality. Generating a list of heuristics for any given unit of instruction would be a major accomplishment but well worth the effort for the clarity it would provide. Using current instructional and assessment protocols could augment current curriculum descriptions as a starting point. Subsequent debates among stakeholders about amendments to the heuristics, their scope and language, can proceed with maximum information about exactly what is recommended and why.

Developing the heuristics should represent a consensus of educators and other relevant contributors about appropriate goals of education and, therefore, of assessment. The process of developing them can be tedious and even contentious, but once finalized, they become a clear specification of instructional targets for educators and students prepared toward them should thus have been given the opportunity to learn. Without them, a match between the assessment and the curriculum can be haphazard and almost certainly is inadequately defined as instructors and examinees are groping to learn (i.e., guessing) about what is in the minds of test developers who themselves may be operating without sufficiently effective shared guidance. Without adequate specification, the resulting domain groping will almost inevitably produce mismatches between examinee preparation and testing tasks, and those mismatches will likely be uneven across examinee preparation programs, producing an inevitable and insidious preparation bias across examinee groups. Assessment professionals can provide the leadership in development of heuristics for specific testing programs; it is up to curriculum specialists to decide whether and/or how to extend the activity beyond the test domain to the larger curriculum.

Published for all to see, tables that include heuristics can remove much of the secrecy that surrounds assessments (and perhaps remove ambiguity on the part of test developers about what constitutes the domain they are working with). It seems difficult to justify hiding the domain of a test from any group, teachers, students, item writers, or test users. As long as the domain represents the curriculum and is expressed at the level of heuristics, the efforts of everyone can be aligned with what is intended in the curriculum. An extension could be to group the heuristics into a two-way table (content by activity) using some convenient taxonomy of cognition (activity categories, such as in Anderson & Krathwohl, 2001), as in a table of specifications (Fives & DiDonato-Barnes, 2013). That table could be useful in sampling the domain to create aligned forms of the test, as I suggested in Schafer (2011), and the process can produce its own documentation.

With specificity comes useful clarity, but some may feel the attendant visibility invites concerns. One concern is to question the value of specific heuristics, yet that can foster healthy debate and perhaps amendments that can move education in improved directions that are more accepted by the public and educators alike. Another concern is that instruction may be more focused on the narrowed curriculum, but the domain will have been designed to represent a consensus about the most valuable instructional goals (our team in Maryland strove for a test domain that covers about 60% of a full course); teachers (and institutions) who feel their students are more capable can then provide a richer experience. Some will feel that more complete domain descriptions will threaten test security, but since the heuristics only suggest test tasks, no task-specific content is revealed. Further, the link between heuristics and items can be used to document that the test does or does not carefully represent the domain/curriculum that is intended. Finally, if instructors teach a well-designed domain of heuristics, they will be teaching the intended curriculum. Teaching to the test domain is a good thing as long as the domain is worth teaching to.

Companion to domain description is description of a sampling plan. Limitations on test length obviously require less-than-complete sampling of the domain. How the domain is sampled should be described and match the test to the plan evaluated for each iteration. It is also possible to incorporate an item development and independent review process that can be used before sampling to yield the needed documentation to document virtually guaranteed alignment in forms over time (Schafer, 2011).

While my comments in this paper are more directed toward educational accountability testing, they could apply to any cognitive assessment, including selection (admissions) tests and competency certification tests, since all these fundamentally ask examinees to do something with something in each task. Thus, there are cognitions and contents that can be (one might say should be) clearly specified in heuristics. It is not too much to ask test developers to specify their assessment domains understandably, especially for high-stakes assessments. Neither would it be too much to ask test developers to justify the heuristics represented by the test program.

Once an elaborated set of heuristics has been developed, it can be used along with sampling decision rules as a test blueprint for selecting and/or developing items to appear on any given form. This process is discussed in Schafer (2011), where it is also suggested that the adequacy of the item pool (e.g., quantity of useable items for each heuristic or combination of heuristics) can be compared with needs within the cells to carry out the sampling and item writers directed to heuristics where tasks are most needed. Another criterion is the richness of the pool of released items. An eventual enhancement could be to add difficulty as a third dimension and adjust the sampling rules so that the other dimensions (content and activity) of the domain are represented similarly for items of low, moderate, and high difficulty and perhaps to assess item-writing needs in order to accomplish equivalent domain sampling algorithmically at all levels in a computer-adaptive environment. These extensions would enhance the quality of the resulting assessments and could become evidence to support the assessment program's validity.

Contextualizing Assessment Results

Test scores are actually quite useless without contextualization for interpretation. Contexts fall into two fundamental types: norm-referenced and criterion-referenced. Norm-referencing involves comparing scores with scores of others in defined groups. Criterion-referencing can be harder to define. On the one hand, the criteria may be the tasks an examinee can do at each given score level (e.g., proficiency level descriptions). On the

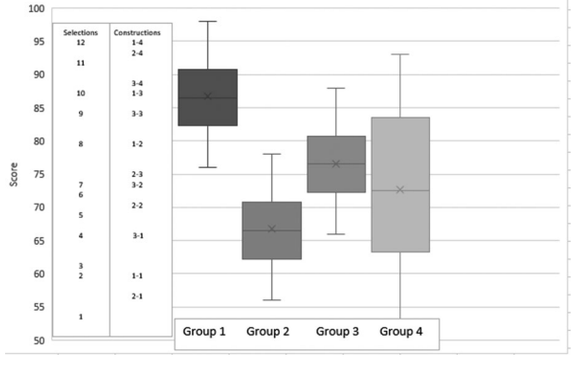
other hand, the criteria may be thought of as achievement levels (normally using standard-setting studies), perhaps with or even defined by consequences (e.g., pass or fail decisions for certification of competency). Depending on the user, any of these ways to contextualize may be most meaningful. Thus, we should anticipate a need for a flexible test-interpretation tool that can be modified by users to yield the most pertinent information to address questions they may have of the data. I will describe a possible approach to meet these goals using a web-based device for attaching meaning to assessment outcomes.

A key difference between what is to be described and current practice is that users are given the means to ask their own questions of the data. At present, officials connected to the testing program typically decide what questions the stakeholders want to (or perhaps should) learn about. But, if individuals had control over the data to be presented, they could focus on their own needs. From the point of view of the testing program, this means large masses of data need to be reconfigured in real-time, a potentially prohibitive task. The compromise described below relies on two tractable sorts of information, item difficulty and group performance, with two types of scales, item scales and data scales, both as the user requests, alongside a scale presentation for the test calibration.

Figure 1 shows a portion of the possible output from a menu-driven website where item columns (selections and constructions), as well as boxplots (Tukey, 1977), appear on a score scale (the author thanks Catherine Eylem for constructing this example). The user can choose the test (giving the scale), whether or not the item columns appear, and which groups to graph. There may be any number of groups plotted, up to the capabilities of the site developer. They could be drawn from an archive of publicly available data, all calibrated on the scale of some test that the user wants to interpret, which itself is selected from a menu of available tests from a maintaining organization. Ideally, the archive would have historical results for a wide variety of relevant groups of examinees. The results needed per group are only five: the 5th percentile, the 25th percentile, the 50th percentile, the 75th percentile, and the 95th percentile in order to populate a boxplot without the outliers. The site also allows the user to input those five results (which could be included routinely in assessment reports) for his or her own group(s). The boxplot(s) are graphed on the scale of the test, say in a column format. Box plots can be constructed by the software in real-time, so only five data points are needed to store to offer them in menus for any groups the assessment would like to facilitate (the more, the better from the point of view of the site user). Call these data columns.

Figure 1

Sample of an Interpretive Data Screen



The item columns should contain links to publicly released items, with the items labeled by type, selected-response or constructed-response. The links appear along the calibration scale and are located by their RP67 values. For selected-response (S), RP67 is the scale value where the item model predicts two-thirds of examinees will answer correctly; for constructed-response (C), each credit-awarded score point has a different location, which is where the probability of a score that high or higher is two-thirds. For an S item, clicking on the link produces the item, the key, and a rationale for the key. For a C item, the link produces the item, a model response for the item at that score level (e.g., 2-3 would be a score of 3 on item 2), and rationales for why not lower and why not higher; if the item is passage-dependent, the passage is also included, perhaps with its reading-level, such as "fifth grade." A nice enhancement would be also to identify the heuristic(s) assessed by the item. The densities of items along the scale can suggest to test developers at which difficulty levels more public-release items are needed and for which heuristics, with item-writing assignments as appropriate.

Examinees could locate their own results on the scale for interpretations of their scores, as locations in the box plots or as the items they would be expected to answer (along with the tasks they would need to be able to perform at higher levels). Instructors and administrators could generate comparisons of local group box-plots with larger groups selected from the menus. Trends over time can be displayed by choosing people columns that differ only by consecutive years. If there are performance levels defined by score ranges, horizontal lines could appear across the columns as desired by the user.

The scale in this example is taken from Schafer and Hou (2011), which includes a process for moderating assessment scales horizontally (across content areas) as well as vertically. The process uses splines to achieve a system that incorporates panel recommendations with normative information for consistent interpretations across content areas (but other methods, such as best-fitting logistics, could be considered for the same purpose). This could expand the interpretations from the website to comparisons of strengths and weaknesses for individuals and groups.

Such a tool could be used by students and parents to interpret assessment scores against groups of students, to see what an assessment score suggests an examinee should be able to do as well as what examinees at higher achievement levels can accomplish, and to compare their school with others. Teachers, too, could compare their students' scores with those of others, compare their students' performance over time, and compare performances on subtests. Principals could evaluate their teachers' results against each other as well as other schools. Policymakers could use the graphics to compare school and district performance levels for resource allocation. Parents might use the information to justify needed improvements in their students' schools.

An attractive feature of this tool is its flexibility. The user can choose any number of columns to view, up to a maximum, of course, for example, seven. The user can graph the same test over time in people columns. Assuming the same scale is used, performances can be compared across subtests. Given the five percentiles for each, local groups can also be plotted to make comparisons with groups defined by the assessment program-supported menus, to make comparisons among local groups, or to interpret individual scores using local norms.

This example approach is designed to be as open as possible in order to satisfy multiple users. The website contains only publicly available information and could be open-access (no password protection). It would be associated with a standardized testing series where test iterations (forms, e.g., over time) are scaled to a common calibration. Institutions may contribute data to the site as long as they are consistently calibrated.

Conclusion

If, indeed, mystery exists surrounding what tests measure, how they measure, and what the results mean, then devices like those described here should help eliminate that source of public distrust of tests. Beyond that, there is much that can be done to explain how content experts, policy advocates, and psychometricians could work together to produce meaningful and useful assessments of important achievements that are as valid as possible. But none of that will happen if our processes for test development remain mysterious and if users cannot ask questions of the data that are meaningful to them. It seems incumbent on test developers to use devices like those described here (and/or others) that clarify and document what assessments actually are in terms that can be understood by the public, whose acceptance is necessary for continued and perhaps expanded use of cognitive assessments. No one else can do that!

References

Albano, A. D. (2021). Commentary: Social responsibility in college admissions requires a reimagining of standardized testing. *Educational Measurement: Issues and Practice*, *40*(4), 49–52.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing*. Longman.

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research, and Evaluation*, *18*, Article 13. https://scholarworks.umass.edu/pare/vol18/iss1/3

Koljatic, M., Silva, M., & Sireci, S. G. (2021). College admissions tests and social responsibility. *Educational Measurement: Issues and Practice*, *40*(4), 22–27.

Schafer, W. D. (2011). Aligned by design: A process for systematic alignment of assessments to educational domains. In G. Schraw & D. R. Robinson (Eds.), *Assessment of higher order thinking skills*. Information Age Publishing.

Schafer, W. D., & Hou, X. (2011). Test score reporting referenced to doubly-moderated cut scores using splines. *Practical Assessment, Research, and Evaluation*, *16*, Article 13. https://scholarworks.umass.edu/pare/vol16/iss1/13

Schafer, W. D., & Moody, M. (2004). Designing accountability assessments for teaching. *Practical Assessment, Research, and Evaluation*, *9*, Article 14. https://scholarworks.umass.edu/pare/vol9/iss1/14

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.