JOURNAL OF APPLIED MEASUREMENT, *24*(1/2), 9-13

COMMENTARIES

A Review of *Clarifying Inputs and Outputs of Cognitive Assessments*

Peter Behuniak
Criterion Consulting

This article examines two elements that are important to all educational assessments. The first is the specification of the content domain to be assessed (inputs). The second deals with the methods used to communicate the results of the assessments (outputs). The author discusses each of these elements, describes why they are integral to the assessment process, and offers ideas regarding how these aspects of assessment could be improved. This review will first discuss the author's treatment of these two topics and then conclude with some summary observations worthy of consideration in the field of large-scale assessment.

Specifying the Test Domain

The author's discussion in this article is mostly focused on large-scale educational achievement testing, such as those created by state or federal agencies to be used for census or sampled assessment purposes. Accordingly, most of the comments in this review will be offered in this context. The article does make some references to other types of assessments, such as licensure and college admissions testing. This review will discuss assessment for these different purposes in the final section.

It is certainly true that the author's first premise, clearly specifying the domain on which a test is based, is a critical initial step in the development of an educational assessment. Creating measures that can support content validity when the results are interpreted is an important part of a test developer's responsibility (Anastasi, 1982; Cronbach, 1971; Shepard, 2009). Accordingly, every public agency and private organization responsible for building assessments include components to address this need in their test development procedures. The most common process involves the work of curriculum specialists to establish the content domain to be assessed. For state agencies, this is often the curriculum adopted by the state for each content area and grade. At the national level, the National Assessment of Educational Progress (NAEP) is based on content frameworks as established by the National Assessment Governing Board (National Assessment Governing Board, 2023) for each content area and grade.

Request for reprints should be sent to Peter Behuniak, 29 Fawn Run, Glastonbury, CT 06033, USA; email: peterbehuniak@cox.net

Test developers must frequently treat these manifestations of content as a starting point rather than the final domain specification. This is due to various practical considerations (referred to as "assessment limits" in the article), such as available testing time, financial constraints, or other resource limitations. In these cases, procedures must be implemented to refine and further specify the content domain and make it manageable for assessment purposes. This is part of the process of developing a test blueprint, and it is to this function that the article's suggestions regarding the specification of the test domain apply.

The author proposes that the use of heuristics could improve understanding of the test domain. I agree. In fact, I think it is a necessity to employ some means to accomplish a transparent statement of the content being assessed. While the term "heuristics" may not be in widespread use, most large-scale assessment programs employ some procedures designed to achieve the same goal. For example, some programs use depth-of-knowledge methodologies to ensure that the assessment applications possess sufficient content validity. Employing these approaches helps ensure that the full breadth and depth of the content domain is represented in the final form of the assessment. The suggestion that heuristics may offer a useful tool for this part of the test development process is worth considering.

It would be helpful for the author to further elaborate on the proposal to use heuristics in the domain specification process. Recognizing that not everything can be addressed in one article, readers could benefit from the illustration of how heuristics could increase clarity. Perhaps it would be possible to provide examples of how heuristics could improve the transparency of some less than adequately defined elements of a given domain.

Contextualizing Assessment Results

The article discusses the need for an effective assessment program to be understood by the stakeholders who rely on it. Examples of these stakeholders include students, parents, and teachers. In many cases, this list could be expanded to include educational administrators, policymakers, and the general public. Certainly, the desire to make test results more meaningful and transparent to stakeholders is a worthy ambition and has long been a high priority for test developers (Ysseldyke & Nelson, 2002).

The strategy proposed is to arm those individuals interested in interpreting the results of an assessment with an online tool that allows flexibility for the user. This would extend what a number of large-scale assessment programs have already implemented in their efforts to provide stakeholders with greater access to results. These efforts vary in sophistication, ranging from little more than the provision of online access to digital versions of printed reports to software that allows some flexibility for the user to explore an existing database. So, overall, I find the proposed online tool to be a reasonable option, but I do believe the potential gain in flexibility brings with it a few caveats.

The author states that test developers "decide what questions the stakeholders want to (or perhaps should) learn about" when determining how test results are to be released (Schafer, 2023, p. 6). This is true. However, there is a good reason for this, which is not addressed in the article. It is a test developer's responsibility to support valid test result interpretations. One of the implications of this responsibility is that a test developer must avoid creating reports or reporting systems that could foster invalid interpretations. A potential problem of providing greater flexibility to users is that it lessens the control test developers have in their quest to guard against unwarranted interpretations of assessment results.

As a simple example, consider the discussion in the article regarding users' ability to employ the online tool to make group comparisons. Large-scale assessment programs routinely adopt guidelines that limit subgroup comparisons unless minimum group sizes are obtained. Now, I realize that this required minimum can be implemented with the proposed online tool (if it is not already part of the system). The bigger concern here is not the example regarding group sizes per se but, rather, the fact that this threat to the valid test score interpretation has to be anticipated in advance in order to avoid the problem. The essence of this concern, then, is whether the designers of this online data manipulation system can anticipate other potential threats and take the steps necessary to eliminate them.

Consider another example involving static group comparisons and trends over time. Output, as displayed in Figure 1 (Schafer, 2023, p. 6), employs five data points per group: the 5th, 25th, 50th, 75th, and 95th percentiles. (As an aside, it is not clear how these data can be provided to individual users unless the examinee groupings are predetermined by test developers. This seemingly would require developers to anticipate the groupings desired by users and would limit the flexibility of the system.) While the availability of these percentiles allows for the graphical display of the groups' results, it is insufficient to support statistical comparisons of group differences. Allowing users (who will necessarily differ in their sophistication and knowledge of statistical group comparisons) to draw conclusions based only on their visual examination of graphical assessment results represents another potential threat. Even if this issue could be overcome, it provides an additional example of a potential threat to the valid interpretation of the test results.

Despite these caveats, there are good reasons to still consider the proposed strategy. To the degree that threats to validity such as the ones discussed can be anticipated by test developers, steps can be taken by the system designers to build in safeguards. Also, it is common practice for large-scale assessment programs to utilize trial periods when implementing new procedures. It is possible that many potential threats could be identified and remedied using such an approach.

Other Considerations

One aspect of large-scale assessment that is relevant to contextualizing the suggestions presented in the article involves the different purposes to which these measures are applied. Clearly, test developers have different purposes in mind when they create assessments designed to measure achievement, make licensure decisions, or aid officials in the evaluation of candidates for college admissions. However, even if we restrict the focus to achievement testing, it is typically the case that a variety of purposes apply. Consider the types of assessments mandated by federal regulations and employed by state agencies. These assessments are used simultaneously for multiple purposes, such as monitoring student progress, curriculum evaluation, accountability issues, evaluation of teacher effectiveness, trend analyses, and more. To be sure, all states are not identical in the purposes for which they employ their assessments, but they all use their assessments for more than just one reason (Behuniak & Way, 2022).

The importance of considering the test purpose here is due to the fact that it provides a context in which the suggestions offered in this article could be viewed. Consider the impact of test purpose on domain specification. Let us use the measurement of grade 4 mathematics achievement as an example. The specification of this content domain will differ in large and small ways from one jurisdiction to another. If the purpose of the assessment in a given jurisdiction is primarily to establish achievement trends over time, the demands on the specification of the content domain are rather low. Generally, only two requirements need to be satisfied for the assessment to fulfill its intended purpose. The specifications need to reasonably reflect the implemented curriculum and need to be held constant over the time period in question.

Contrast this with another common assessment purpose, curriculum evaluation. In this case, it is critical that the assessed domain mirrors the existing curriculum. The teachers, curriculum specialists, and others attempting to

use the test results to draw conclusions about the range and depth of the existing curriculum could only do so if the assessment covered the entire curriculum at a grain size that was as close to equivalent as possible. The details associated with the domain specification of the assessment would be of the highest priority and would certainly be subject to great scrutiny by all involved stakeholders.

A similar case could be made for how differing purposes affect the presentation of test results. Staying with the above example, the reporting of assessment results to support a trend analysis would need a different look than if curriculum evaluation was the goal. The trend analysis could be accomplished in many ways ranging from the presentation of simple mean scores to analyses conducted at a much finer level of detail. However, an effective evaluation of the curriculum would necessarily require reporting units that match the way the content is specified in the established curriculum.

The concerns regarding test purpose are discussed here as a way to view the author's suggested approaches in context. The fact that large-scale assessments are used for multiple purposes does not lessen the potential benefit of the approaches suggested in the article. It does mean, however, that the ways in which these suggestions should be best implemented may differ depending on which purpose is being served. It is also conceivable that some purposes might obviate the utility of one or both of the suggestions.

A second consideration that is relevant to viewing these suggested approaches in context involves the stakeholders in the assessment process. In large-scale achievement testing, these include students, teachers, educational administrators, policymakers, media, and the general public. These groups do not view assessments or assessment results in the same way. Since their expectations differ, what they hope to learn when they review test results will also be different. Evaluating the potential benefit of suggestions regarding domain specification and the reporting of results will rely, in part, on taking into account these differences among stakeholder groups.

To illustrate this point, consider the release of NAEP scores. When NAEP results are released to the media and the general public, the focus is usually on the comparisons that the recent results support. Are the scores higher or lower than in previous years? Are boys showing more or less progress than girls? Are the achievement gaps between different demographic groups closing? These types of questions are generally answered using score means, score differences, trend lines, and similar presentations. There is very little attention paid to the specification of the content domain. This is mostly due to the fact that these stakeholders generally accept the fact that knowledgeable professionals put in the time and effort to define the content domains (i.e., NAEP frameworks) appropriately. They would be right.

Compare this to what happens leading up to the release of NAEP scores. The process of establishing the content domains for NAEP takes years and involves input from hundreds, perhaps thousands, of content specialists, including representatives of every major professional organization with an interest in the target content area. Every detail and element of the proposed content domain is examined and discussed until a final framework is adopted. On a smaller scale, this type of process is carried out by state agencies during the development process for state assessments. These content specialists, whether operating at a national or state level, are deeply interested in the details of the domain specification.

Evaluating the potential of the proposed suggestions will be enhanced by taking into account the differing interests of each stakeholder group. The degree to which the suggested approaches improve the transparency of assessments can be increased by considering the level of transparency needed to best satisfy the interests of each group. The consideration of stakeholder needs and desires is routinely taken into account when large-scale assessment programs establish their domain specifications and reporting procedures. These same needs and desires are relevant in evaluating the potential of the suggested approaches as proposed in this article.

References

Anastasi, A. (1982). *Psychological testing* (5th ed.). Macmillan.

Behuniak, P., & Way, D. (2022). *White paper to provide context for NAEP achievement levels by reviewing state and international practices*. American Institutes for Research, NAEP Validity Studies Panel.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

National Assessment Governing Board. (2023). *National assessment governing board assessment framework development policy statement*. https://www.nagb.gov/content/dam/nagb/en/documents/policies/assessment-framework-development.pdf

Schafer, W. D. (2023). Clarifying inputs and outputs of cognitive assessments. *Journal of Applied Measurement*, *24*(1/2), 1–8.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practices*, *28*(3), 32–37.

Ysseldyke, J., & Nelson, J. R. (2002). Reporting results of student performance on large-scale assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy and implementation* (pp. 467–480). Lawrence Erlbaum.