JOURNAL OF APPLIED MEASUREMENT, *24*(1/2), 39-43

REJOINDER

Reactions to the Commentaries on Clarifying Inputs and Outputs of Cognitive Assessments

William D. Schafer
University of Maryland

I want to thank the professionals who took the time to read and react to my thoughts on ways to better communicate with the public about what we assess and how we interpret assessments. I also want to thank the editor of the *Journal of Applied Measurement* for his encouragement to express these ideas in a more formal way and for the opportunity to comment on the reactions in the responses he received, all in our shared goal of improving educational practices through more effective use of assessments.

Perhaps the best way to organize my responses is to comment separately on each of the four reaction papers. Although there is some overlap, which I will note from time to time, each reaction has its own perspective, and that makes it worthy of consideration on its own.

Behuniak (2023)

I should begin by mentioning that Peter Behuniak served Connecticut in a role similar to mine in Maryland (I was Director of Student Assessment), which I took on after almost thirty years as a faculty member in the University of Maryland College of Education, Department of Measurement, Statistics, and Evaluation. Our experiences with mandated assessments of students are thus reasonably similar, though the two states differed significantly at the time in the assessments they administered and how they were used (late 1990s).

Behuniak is quite accurate in pointing out that my focus is on high-stakes assessments mandated by governmental or quasi-governmental agencies. Stakes can be for examinees or for institutions, but clearly less formal events, such as classroom assessments, though very important in student learning, are outside my current purview. Testing for other purposes, such as selecting among candidates and licensing or certification, also seems appropriate for my suggestions, but other considerations may impinge on relevance.

I have suggested a specialized definition of the term “heuristic,” and Behuniak reasonably suggests that elaboration is needed. His reaction is highly appropriate, and I would like to take the bait, so to speak, and expand on that. Perhaps an example based on my experience

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

in Maryland with our work designing course-based, high school assessments required for graduation would be useful. In describing the domain of a new assessment for a government course, one topic (paraphrasing from memory) was to have an understanding of landmark Supreme Court cases. As one teacher coordinator said to me, how can a teacher (or curriculum developer) know which cases are "landmark" and what constitutes an "understanding?" Our assessment team discussed that point and decided to identify ten cases as "landmark" and elaborate on what knowledge is needed for each (e.g., context, disputed issue, arguments on each side, the court's decision and reasons, and implications for society). The government content coordinator gathered about 20 government specialists from around the state, and a (very!) spirited discussion finally resulted in a consensus that gave educators clarity about the testable features of landmark Supreme Court cases. Our assessment teams in each of the four tested areas (algebra, biology, government, and language arts) developed similar statements that exhibit the characteristics of heuristics, expressing agreed-upon assessment limits.

Regarding the interpretation of the graphic device I suggested, Behuniak points out that techniques can be used inappropriately, and that is especially true of assessments. I agree and have found over the years that anticipating all possible unintended consequences is both important and hopeless. I can only suggest that the study of uses be incorporated into implementations of this (and any other) interpretation tools used in practice. A helpful by-product of such work could be to suggest additional ways to encourage positive uses.

A note is needed about constructing the graphs in Figure 1 of the paper. I presume each assessment series (e.g., government tests) is the product of an agency that can easily calculate and store the five percentiles for all takers and for each appropriately sized (perhaps 40 or more) demographic group and for sub-jurisdictions down to the building level. These data points could be accessed from menus by any user and boxplots developed by the software. Anyone who has access to data on lower levels, such as building principals or content supervisors, could calculate the percentiles for their own smaller groups rather easily through simple counts.

I feel that the use of graphical comparisons of group differences can help avoid overinterpretation of group differences, a common threat in releasing only central tendency measures. In the case of the graphs, it should be relatively clear when minor differences in location are swamped by the variability present in the results for each group.

Hau, Xiao, and Guo (2023)

The reactions of Hau et al. and Tseng are especially interesting, coming from cultures very different from my own. Nevertheless, I am struck by the consistency of goals and concerns in the area of assessments and education in general.

Hau et al. appropriately differentiate low-stakes and high-stakes assessments, and individual student and group levels. Clearly, my concern with the interpretation device was for high-stakes assessments, with the stakes existing for either students or institutions. I am envisioning a program that issues examinations over time using a unidimensional IRT model to place items and students on the same axis (though extensions to a multidimensional context might be studied). They questioned whether the suggestion to amalgamate reports into one device is appropriate. Actually, it never crossed my mind that current reports might be supplanted by the device. I think those that would become unused would be discontinued as users found them less helpful in practice.

Considerations were raised by Hau et al. (and Tseng) about an artificial narrowing of the curriculum to a limited range of content topics as well as to lower-order cognitive processes. Each of these deserves consideration.

In its development of course-based high-school assessments, Maryland faced the concern about narrowing the content over which students received instruction. The following approach seemed to work quite well. In determining the breadth of each examination, the curriculum committees established what they considered to be the content anyone should expect a student taking that course to exit with. It was decided to limit the domain of each test over the content that could be included within 60% of the material in a solid course, leaving the other 40% to the discretion of the district, the school, and/or the teacher. The content not included in the state's domain could be assessed locally if desired.

The concept of heuristics was partially intended to address the concern of too much emphasis on lower-order, rote memorization. Recall that a heuristic needs to be broad enough to allow multiple assessment opportunities as well as narrow enough to be able to apply it in determining whether given talks are or are not within it. The verbs used can be those that generate higher-order cognitions. Such terms as clarifying, paraphrasing, illustrating, generalizing, applying, outlining, structuring, testing, hypothesizing, and explaining are examples. The heuristics, themselves, can be evaluated for cognitive demand, but only if they are stated rather than implied.

My feeling is that any high-stakes testing program will, of necessity, narrow the curriculum. With explicit heuristics, the narrowed subset will be understood and agreed upon by the most salient stakeholders. If they are not agreed upon, let alone explicit, there will be guessing on the part of educators about the scope of the assessment, and these guesses will lead to preparation programs with somewhat haphazard goals. This is a source of invalidity from the perspective of assessment developers and of unfairness on the part of examinees. Could high-stakes assessments written over non-explicit domains even be seen as testing malpractice (I think so, but other professionals may not express it that strongly or may disagree completely)?

Hau et al. raise an interesting point about scoring rubrics. Especially in the area of creativity, they suggest students could exploit a rubric to achieve a higher score than they deserve. In any instance where it occurs, that may be the fault of the rubric or perhaps the prompt the examinee is responding to. Hopefully, the assessment is designed to be valid, where higher scores represent more of the trait being assessed. It is hard to comment without a specific example, but I am reminded of a facetious essay written by a Maryland scoring staff member to illustrate that a high score could be achieved through automated scoring for a logically worthless essay written to conform to sound principles. He was successful. But one could argue that the text was written by someone who does indeed deserve a high score even though the text itself does not. I am trying to illustrate the point that one may be able to "trick" a rubric, but perhaps that itself is evidence of achievement.

Tseng (2023)

With good reason, Tseng raises some practical concerns about implementing the suggestions in my paper. One has to do with linking tested domains with curricula across political units. What is taught in states differs considerably, and differences across countries are more profound. This is clearly true, and in many states, there even seems to be a political goal of emphasizing these differences. This phenomenon suggests that content differences are an important and perhaps crucial element to consider in making comparisons between the units. I would hope that explicit statements of heuristics would help make the differences more explicit and, therefore, easier to debate, discuss, and perhaps move toward a consensus about a valid assessment system for purposes of comparison.

Comparisons over time do indeed require consistent assessments. One often hears the admonition, if you want to measure change, don't change the measure. But curriculum does change, and as Tseng (and Behuniak)

mentions, assessed domains need to represent current curricula. This is an issue faced by virtually every dynamic jurisdiction, and several ways to address it have been proposed. None satisfy all possible criteria. As a workable example, consider a change in curriculum where some material is deleted and other is added, presumably with its own heuristics. One might introduce the change first in curriculum and instructional materials, and in the next year, by items in assessments used only for item evaluations and tentative calibrations, and in the next year by scored items that are post-equated, and in the fourth year by items whose calibrations pre- and post-calibrations are consistent. Finally, released items on the new content can be added to the interpretation system as available.

While not my focus in the current effort, I want to support strongly Tseng's comments on formative assessments and the need for assessment literacy on the part of teachers. These are issues quite meaningful to me, having written on both assessment literacy in educational programs for various school roles and on formative assessments, as well as founding the Classroom Assessment Special Interest Group of the American Educational Research Association.

The goal of general assessment literacy on the part of the public is difficult to achieve, if not Sisyphean. I have tried to use the interpretive device only techniques that are simple and straightforward. Boxplots, for example, are often included in elementary school curricula.

Bezruczko (2023)

Bezruczko embeds the current zeitgeist surrounding educational tests in the history of testing and calls for a fundamental re-thinking of assessments in their social context. He suggests needs exist that more broadly define traits that should be assessed and would be supported by the public. Perhaps and perhaps not. His suggestions are well-reasoned but debatable. In the end, though, the devil is in the details. Can traits that are more personal (non-cognitive) attributes be assessed effectively enough to make decisions about people? Should these be seen as outcomes of schooling? Should tests be used to identify individual differences? These are examples of the sorts of questions that require public debate but, in the end, are tangential to the problems I tried to address in my paper. I hope some specific suggestions can appear that give meaning to work in the directions he suggests.

Final Thoughts

The two thrusts I suggest can be thought of as complementary. If heuristics are slotted into blueprints, it will be possible to isolate cells that are underrepresented by items. Very specific directions can then be given to item writers to generate needed coverage. Interestingly, it will also be possible to generate items across difficulty levels in each of the cells of the blueprint.

The pool of released items can be evaluated in the same way. Subtest coverage of content and cognition can be enhanced so that users of the system of boxplots and item maps can be fully informed about what constitutes performance across the spectrums of content, cognition, and performance.

Finally, please note that all information, from the heuristics to the blueprints and the released items to the five percentiles for all demographic groups maintained by the assessing institution, can (or should) be publicly available. The graphical device recommended here is my suggestion to package already public information to best meet the needs of a broad range of users, helping them to reach appropriate, supportable, and useful conclusions. And that is, I believe, a goal all of us share.

References

Behuniak, P. (2023). A review of clarifying inputs and outputs of cognitive assessments. *Journal of Applied Measurement*, *24*(1/2), 9–13.

Bezruczko, N. (2023). Reactions to "Clarifying the inputs and outputs of cognitive assessments." *Journal of Applied Measurement*, *24*(1/2), 23–38.

Hau, K.-T., Xiao, L., & Guo, L. (2023). Inputs and outputs of cognitive assessment: Navigating the complexities of multiple purposes and end-users. *Journal of Applied Measurement*, *24*(1/2), 14–18.

Tseng. F.-L. (2023). A commentary on clarifying inputs and outputs of cognitive assessments. *Journal of Applied Measurement*, *24*(1/2), 19–22.