JOURNAL OF APPLIED MEASUREMENT, *24*(1/2), 14-18

COMMENTARIES

Inputs and Outputs of Cognitive Assessment: Navigating the Complexities of Multiple Purposes and End-Users

Kit-Tai Hau
The Chinese University of Hong Kong

Leifeng Xiao
Shanghai Normal University

Luyang Guo
University of Macau

Schafer (2023) championed enhanced clarity in cognitive assessment domains and a comprehensive reporting system. While we resonate with these principles, we underscore the significance of less tangible, evolving higher-order abilities and innovative proficiencies. Furthermore, the objectives behind cognitive assessments and their reports vary across student, school, district, and country levels. Therefore, consolidating these diverse reports into a single system might not be ideal and optimal.

Requests for reprints should be sent to Kit-Tai Hau, Faculty of Education, The Chinese University of Hong Kong, Shtain, N.T., Hong Kong; email: kthau@cuhk.edu.hk

The debate concerning the effectiveness, utility, and fairness of cognitive assessments, especially high-stakes ones, has been a prominent topic in educational discourse. Schafer (2023) illuminated several of the main challenges and proposed methods to enhance the utility and acceptance of these assessments among the public.

Koljatic et al. (2021), along with other commentaries in the same issue of their article, delved deep into the advantages and disadvantages of high-stakes assessments, especially in the context of university entrance examinations. Extending this conversation, Schafer (2023) explored a wider array of cognitive assessments, proposing enhancements to their scope, mode of delivery, and public acceptance. His primary focus was on the high-stakes educational accountability applicable to schools and districts. Nonetheless, he was confident that his arguments had broader implications. This article explores Schafer's propositions in the context of cognitive assessments used at the student, school district, and national levels.

Schafer (2023) began by emphasizing two points: (a) the importance of aligning the domains of major standardized tests with curricular content and (b) the potential benefits of utilizing interactive tools to interpret test outcomes. He delineated the relationships among (a) the curriculum, which represents a subset of the entire body of knowledge and skills; (b) the test domain – a fraction of the curriculum; and (c) the test itself – an operational representation of the domain. Schafer noted the public's discontent with the vague nature of assessment content, summarising the sentiment as: "We are going to test you, but we won't specify what the test will cover."

While we concur with Schafer's general direction, we find it essential to distinguish between two pivotal dimensions of assessments: the distinction between low- and high-stakes assessments and the distinction between assessments at the student level versus those at the school, district, or country levels. In theory, once a large cohort of students is evaluated, we can generate reports tailored to individual students, schools, school districts, and even states or countries. The question arises: Is amalgamating all these reports within a singular system optimal?

Transparency and Clarification of Domain Coverage

Balancing Transparency and Narrowing Curriculum

A pressing concern is the balance between providing a comprehensive examination syllabus, as Schafer (2023) advocated, and the potential drawback of this leading to a narrower curriculum and increased rote memorization. In high-stakes examinations for students or schools (e.g., college entrance test or school accountability assessment) and large-scale surveys (as in national or international surveys), some assessment authorities provide exhaustive examination syllabi, but some do not. For instance, the Trends in International Mathematics and Science Study (TIMSS; Mullis & Martin, 2017) offers clear examination curriculum coverage, whereas the Program for International Student Assessment (PISA; Organisation for Economic Cooperation and Development [OECD], 2019) does not. Schafer (2023) argued that a curricular focus would not be problematic if there could be public consensus regarding key instructional objectives. We concur that a consensus might bolster public acceptance of tests, but it does not necessarily mitigate the issue of curriculum narrowing.

Several related challenges arise. Formal school curricula struggle to encapsulate and quantitatively assess all important educational content and competencies. This challenge is exemplified by PISA's introduction of an innovative domain (e.g., creative thinking in 2022) to incorporate important educational areas beyond traditional subjects. However, expecting such innovative domains to seamlessly integrate

into curricula and provide a comprehensive framework remains a tall order. Schafer's (2023) push for detailed examination curricula might inadvertently hinder the inclusion of challenging-to-assess higher-order thinking skills and innovative domain competencies.

Narrowing Curriculum and Goodhart's Law

Moreover, the potential narrowing of the curriculum is also linked to Charles Goodhart's Law, elegantly put by Strathern (1997, p. 308): "When a measure becomes a target, it ceases to be a good measure." Some assessment tasks and measures, although initially valuable indicators of students' accomplishments, can lose their efficacy when they become the focal point of high-stakes tests. This happens when students employ shortcuts to achieve high scores. Furthermore, students' language performance, for example, can be a good measure of the quality of education in schools. However, when language is chosen as a high-stakes measure, schools may focus on language teaching and ignore other subjects. Then, language performance ceases to be a good measure.

In an ideal scenario, a comprehensive assessment would encapsulate all intricate higher-order thinking skills and innovative domain competencies. The assessment of creativity, for example, is of paramount importance. However, if integrated into high-stakes evaluations, students might exploit the nuances of the scoring rubrics, aiming for high scores without genuinely enhancing their creative abilities. Our use of ideal learning tasks for high-stakes purposes put all of us into an impossibly difficult-to-solve situation.

Evolving Important Competencies

The challenges are not confined to the current curriculum. We must also consider "future" competencies that will be crucial for students as they transition from school to the workforce. For instance, teaching students to drive might become obsolete with the advent of AI-driven vehicles. As such, our assessment domains must continually evolve. With AI writing software becoming mainstream in early 2023, certain skills, like drafting and writing, might diminish significantly. On the other hand, the ability to critically evaluate AI-generated content becomes imperative. This "evaluation" skill ranks at the pinnacle of Bloom's taxonomy.

The quandary of whether or not to provide detailed test specifications might also mirror the broader educational debate: Should we prioritize (a) precise assessment accuracy or (b) the emphasis on hard-to-measure and continually evolving competencies? Advocates of the former might lean towards comprehensively detailing a restricted set of learning objectives to ensure precise student assessment and comparisons. In contrast, proponents of the latter might feel confined by the current assessment paradigms and opt for a more expansive approach.

For high-stakes individual assessments, test authorities might need to define examination content more clearly, hence limiting their assessment of higher-order skills. However, for school accountability measures or international evaluations comparing educational systems, a more open-ended approach, devoid of a stringent behavioral-objective framework, might be more educationally enriching. This latter approach seems more promising given the challenges in defining our assessment frameworks, let alone the complexities of emerging core competencies.

Reporting Assessment: How Detailed and Comprehensive Should it be to Make it Meaningful?

Schafer (2023) distinguished between norm- and criterion-referenced evaluations and proposed a digital system to present the most relevant data for a specific assessment context. He even showcased potential result displays and hinted at the possibility of connecting these results to publicly available items for learning purposes. We agree with Schafer's fundamental philosophy that distinct reports should cater to different user tiers (students, schools, districts, and countries).

Student Assessment and Reporting

In theory, once a test is administered, we can generate reports for individual students, schools, districts, and even countries. However, given their distinct objectives, consolidating these diverse assessment levels into a unified system might not yield optimal results.

For the utmost benefit of students, a system should facilitate regular assessments throughout and following students' learning of a particular topic. As the emphasis is on "assessment as learning," students might be permitted to retake certain segments of the assessment. This approach would necessitate extensive and thorough evaluations of most learning objectives to provide valuable diagnostic feedback. Such assessments are typically low-stakes, with students taking the lead in monitoring their own progress. Oftentimes, we do not even know whether they are the students, their parents, or the elder siblings doing the tests. Thus, aggregating such results to reflect school performance might not be feasible. That is, these results are not suitable for school accountability use.

School, District, Country Level Assessment and Reporting

While we offer in-depth diagnostic feedback for individual student reports, the public often expects similarly detailed reports for schools, districts, or national assessments. The public's interest often lies in discerning whether a specific institution, district, or country is underperforming in a particular academic domain. It is imperative to consider whether an exhaustive domain-by-domain comparison is warranted.

Analyses showed that schools (countries) performed consistently across different academic subjects. That is, a school (or country) that is good (or bad) in one academic subject will also be good (or bad) in other subjects. Based on the PISA 2018 data, there were markedly high correlations and consistency in performance across subjects: the correlations between science, reading, and math ranged from r = .72 to .80 at the student level, .92 to .95 at the school level, and an impressive .95 to .98 at the country level (Hau & von Davier, 2023). This suggests that a student, school, or economy that excels (or struggles) in one subject typically mirrors that performance in others.

Several insights can be drawn from these findings:

1. When ranking schools, districts, or countries, selecting a single academic subject could serve as a reliable indicator. A potential drawback to this is that once a subject becomes the focus, schools might prioritize that subject to the detriment of others. Evaluating all key academic subjects is recommended to counteract such behavior, even if the rankings derived from different subjects are closely aligned and consistent.
2. Granular diagnostic reports at the district or country level might reflect more variability than genuine differences.
3. Obviously, weekly assessments of schools and districts are unnecessary. Evaluations every couple of years could offer fairly consistent results unless school management, teaching force, or student in-take change substantially in the schools or districts.
4. Within such an assessment framework for schools (or districts), having a sample of students undergo brief tests on a range of academic areas (akin to PISA) could yield reliable indicators of a school's or district's quality. However, if there is a need to determine the added value, the assessment design would be more intricate.

For comparisons at the school, district, or country level, besides academic evaluations, it would be beneficial to incorporate essential questionnaires designed to elucidate the disparities among these entities. Rather than simply determining who performs best in these cognitive assessments, it is much more important that we understand the underlying reasons for such performance. However, these

questionnaires should not be used for high-stakes decision-making because the integrity of responses on such tools is hard to guarantee. Students, schools, and districts can cheat easily on the questionnaires. However, it is crucial to incorporate significant policy-related queries to inform system improvements.

Theoretically, a comprehensive system that encompasses assessment and reporting for students, schools, districts, and countries is achievable. However, given the variances in intent, frequency, and report types, it may not be practical to design a single, all-encompassing system to cater to these diverse needs.

Conclusion

While Schafer's (2023) endeavor to enhance the clarity of cognitive tests and make their results more understandable is commendable, a comprehensive elucidation of test content and results might not directly address many of the existing challenges. Meticulous detailing of the assessment criteria might inadvertently impede the incorporation of higher-order thinking skills and emerging essential domains into the curriculum and assessment. As for reporting, while providing a wide array of feedback across different levels is theoretically possible, funneling all these into a singular system might not be beneficial due to the divergent objectives of different assessment levels.

References

Hau, K. T., & von Davier, M. (July 4–6, 2023). *Large-scale international education surveys: Analyses of slopes can be more interesting than comparisons of means* [Paper presentation]. The International Society for Data Science and Analytics Annual Meeting, Fudan University, Shanghai, China.

Koljatic, M., Silva, M., & Sireci, S. G. (2021). College admissions tests and social responsibility. *Educational Measurement: Issues and Practice*, *40*(4), 22–27. https://doi.org/10.1111/emip.12425

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Boston College, TIMSS & PIRLS International Study Center.

Organisation for Economic Cooperation and Development. (2019). *PISA 2018 results (Volume I): What students know and can do*. PISA, OECD Publishing. https://doi.org/10.1787/5f07c754-en

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

Strathern, M. (1997). 'Improving ratings': Audit in the British University system. *European Review*, *5*(3), 305–321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4