The Quest for the Holy Grail of Validity in Science Assessments – A Comment on Kampa and Köller (2016) "German National Proficiency Scales in Biology: Internal Structure, Relations to General Cognitive Abilities and Verbal Skills"

Ronny Scherer

University of Oslo, Norway

Author Note

Ronny Scherer, Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Norway.

Acknowledgements. The author would like to thank the reviewers and the journal's editors-in-chief for their constructive and elaborate comments on previous versions of this paper.

Correspondence concerning this article should be addressed to Ronny Scherer, University of Oslo, Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, Postbox 1161 Blindern, N-0318 Oslo. Phone: +47 228-444 02, E-Mail: ronny.scherer@cemo.uio.no

Keywords: Differential item functioning; Instructional validity; Science assessments; Scientific inquiry; Validity argument

The Quest for the Holy Grail of Validity in Science Assessments – A Comment on Kampa and Köller (2016) "German National Proficiency Scales in Biology: Internal Structure, Relations to General Cognitive Abilities and Verbal Skills"

Introduction

As Duckworth and Yeager (2015) have put it, measurement matters. In fact, progress in science education largely depends on the availability of reliable and valid measures of either highly complex constructs such as scientific inquiry or more well-defined constructs such as the knowledge of facts and content (Bauer, 2016). For a measure to be valid, effective test design needs to fully attend to validity issues. This dependence clearly necessitates the validity of science assessments or, to be more precise, the quality of the inferences, claims, and decisions drawn from the resultant assessment scores (Zumbo, 2006). Even further, the search for validity evidence has somehow become a quest for the 'holy grail', and the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) clearly put forth that the provision of validity evidence is critical for the development of any assessment (Liu, 2010). Part of this reasoning is that reporting validity evidence not only helps those who want to use assessments to understand and interpret assessment outcomes appropriately, it is also part of the scientific process in a sense that it facilitates transparent communication about how well assessments measure what they intended. At the same time, there is no strict dichotomy in reporting validity (i.e., 'validity is given or it is not'); instead, there are only degrees to which an argument for validity can be crafted, and different sources of evidence may support this argument (AERA et al., 2014; Reeves & Marbach-Ad, 2016).

In their recently published paper, Kampa and Köller (2016) present an assessment of students' content knowledge and scientific inquiry skills in the domain of biology and describe the validation thereof. This study showcases the creation of a validity argument of science assessments by focusing on the internal structure and the relations to other constructs,

namely general cognitive abilities and verbal skills. The authors profoundly illustrate how they obtained validity evidence, pointing to potential ways to approach validity from an item and test performance perspective. But what can be learned from this perspective? Which approaches to validity can researchers and test developers in the field of science education take in order to establish or improve their validity argument? Which sources of evidence can inform the quest for validity? This essay briefly reviews Kampa's and Köller's (2016) validity argument and extends on potential ways to supplement it. Along these lines, it sketches approaches to validity science educators may want to take in order to examine different sources of evidence. My comments will be based on the validity framework for instructionally relevant assessments proposed by Pellegrino, DiBello, and Goldman (2016).

Creating a validity argument based on item and test performance

Indeed, there are many roads researchers can take to craft a validity argument. The validity argument stands and falls with the quality and the breadth of its sources of evidence (Kane, 2013; Wilson, 2004). Attempting to specify the approaches to validity, the *Standards for Educational and Psychological Testing* (AERA et al., 2014) include quality criteria and potential sources one could draw from: test content, response processes, internal structure, relations to other variables, and consequences of testing. Pellegrino et al. (2016) approached validity from an instructional perspective organized as cognitive, instructional, and inferential validity (p. 62): *Cognitive validity* refers to the extent to which the assessment captures the relevant aspects of knowledge and skills in ways that are not confounded with other cognitive processes or skills. *Instructional validity* refers to the extent to which the assessment is aligned with curriculum, instruction, and students' opportunities to learn; it also includes instructionally relevant information about how the assessment might support teaching practices. *Inferential validity* refers to the extent to which diagnostic and model-based information about students can be obtained from the assessment. Pellegrino et al. (2016)

propose further evidence to support these three validity facets. Table 1 provides an overview of these sources for studies that primarily focus on studying item and test performance. On the basis of these sources and facets of validity, I review the evidence Kampa and Köller (2016) tap in their paper for crafting a validity argument.

Insert Table 1 about here

Internal structure. The authors examine the internal structure of the national proficiency scales by testing specific assumptions about the relation between students' content knowledge and scientific inquiry skills. Performing elaborate statistical analyses based on uni- and multidimensional item response theory models, they find that both constructs can be empirically distinguished, although they are highly correlated. It is noteworthy that Kampa and Köller (2016) did not attempt to study the internal structure of the two subscales by differentiating between, for instance, different processes and skills of scientific inquiry or different types of content knowledge (e.g., with respect to different scientific contexts or concepts); instead, they tried to disentangle the overall relation between content knowledge and scientific inquiry skills. Although a further differentiation of the two concepts may provide more fine-grained information on the relations among the sub-processes and skills and therefore reveal instructionally relevant information about the specific strengths and weaknesses of performance or conceptual understanding (Hartig & Höhler, 2009; Leighton & Gierl, 2007; Wind & Gale, 2015), the reported findings nicely serve the purpose insofar that evidence for the empirical distinction and relation is provided. At the same time, it must be noted that results along these lines may not be as clear-cut as in the present study. In other situations, researchers may have to accept that constructs or facets of a construct are not

clearly distinct, for instance due to high correlations among the constructs or facets or the occurrence of an unexpected number of factors that are extracted from item responses. Even further, although theoretical assumptions about the structure of a construct may suggest multidimensionality, researchers may not be able to find support for the hypothesized internal structure. Such situations may require suspending the expectations on an assessment's ability to discriminate between factors or constructs and – as a potential consequence – the refinement of previous assumptions on the internal structure that have guided the creation of a validity argument. In this respect, replication studies could help to clarify conflicting results (Duncan, Engel, Claessens, & Dowsett, 2014). Overall, I agree with the authors that examining the internal structure of the assessments is essential for crafting a validity argument, and applaud them for providing reasonable evidence that supports this argument.

From a substantive point of view, the question about the role of content knowledge for scientific inquiry skills has been discussed controversially in science education, particularly with respect to the questions of how much and which type of content knowledge is needed to succeed in inquiry tasks (Kuhn, Iordanou, Pease, & Wirkala, 2008; Leighton & Gierl, 2007; Williams, Ma, Prejean, Ford, & Lai, 2007). Hence, Kampa and Köller (2016) address a critical issue and provide some more insights into the matter on the basis of a sufficiently large student sample and reasonably well-functioning assessments. These insights mainly tap what Pellegrino et al. (2016) called inferential validity (i.e., dimensionality of the assessment, model-data fit) and cognitive validity (i.e., empirical distinction of cognitively different constructs; see Table 1).

Another perspective that extends arguments supporting the validity of science assessments refers to the investigation of group differences, for instance with respect to students' gender, socioeconomic status, or ethnicity. Of course, group differences are by no means trivial, as they not only describe gaps in science *performance* (e.g., Lee & Burkam,

1996; OECD, 2015; Wang & Degol, 2016), but also the quality or, more precisely, the *fairness* of the underlying assessments (Zwick, 2012). From a validity perspective, any performance differences identified may not necessarily reflect actual differences in the underlying construct; performance differences may occur because the assessment operates differently within the groups. Situations in which this is the case point to *differential item functioning (DIF)* – a concept that has found its way into most validity frameworks (AERA et al., 2014; Pellegrino et al., 2016; Wilson, 2004). In this respect, "fairness" represents what is often called *measurement invariance* – that is, "the situation in which a scale or construct provides the same results across several different samples or populations" (AERA et al., 2014, p. 211). If a sufficient degree of measurement invariance holds, valid group comparisons can be conducted and test fairness with respect to these groups is by and large ensured (Millsap, 2011). Once again, because the investigation of measurement invariance focuses on the extent to which the measurement model and its statistical parameters are comparable across groups of students, researchers can make inferences on the functioning of an entire test or specific items. The authors of the current study discuss this issue (Kampa & Köller, 2016) and point to the need for examining whether or not the relation between content knowledge and scientific inquiry is subject to differences across educationally relevant groups. These groups may be defined by categorical variables such as gender, ethnicity, socioeconomic status, educational track, and nationality (Babiar, 2010; Eggert & Bögeholz, 2010; Wilson, 2004) or continuous variables such as verbal skills or general cognitive abilities. For instance, science assessment may work differently for students of different ethnicity, perhaps due to the fact that students understand specific items differently given their cultural or language background. In this respect, the detection of DIF helps test developers to understand how assessments work in specific groups of students and to possibly refine these assessments to achieve comparability. Still, to draw conclusions on how assessments should be refined, researchers should strive for

explaining why DIF occurs. Qualitative data collection activities might accomplish this. Overall, an indication of DIF helps researchers and teachers who may want to use the assessments in future investigations or in the classroom to interpret the scores from a more differentiated perspective; it may also inform test developers about the necessity for group-specific items, scales, or tests. I believe findings along these lines could significantly strengthen researchers' argumentation on the inferential validity of assessments.

Relations to other constructs. Kampa and Köller (2016) chose to examine the relations between the two science constructs – content knowledge and scientific inquiry skills – and the two cognitive covariates verbal skills and general cognitive abilities. Their findings make the case for differential relations and strengthen the evidence that content knowledge and scientific inquiry may comprise different cognitive processes. This evidence once again addresses the cognitive and inferential validity of the assessments (Table 1). It also reveals that students' performance is to some extent influenced by other competences and points to the fact that no measure is pure and perfect (Duckworth & Yeager, 2015). I argue that the authors' hypotheses are reasonable and the results supporting them somehow expected. For instance, general cognitive abilities have long been considered indicative of students' information processing (Sternberg, 1977). As information processing plays an essential role for most cognitive abilities (Evans & Stanovich, 2013), it is not surprising that it does explain variance in both content knowledge and scientific inquiry skills. What might be surprising to those who believe in the generality of cognitive abilities is that the variance explanation is only about 50% even after controlling for verbal skills. This, in fact, makes the case that the proposed science assessments measure something that is not entirely a composite of general cognitive abilities and verbal skills. Nevertheless, the assessments of these covariates were rather limited and warrant assessing further dimensions to strengthen the case (e.g., by using not only figural but also numerical and verbal dimensions of general cognitive abilities). In

light of this finding, I am convinced that both content knowledge and scientific inquiry in biology go beyond these abilities and skills. Kampa and Köller (2016) consequently provide some evidence on the specificity – yet not the generality – of the two constructs under investigation.

Despite the confirmatory approach Kampa and Köller (2016) have taken, researchers may want to extend the selection of covariates by including relevant "non-cognitive" and beliefs-oriented constructs depending on the construct under investigation. Specifically, the authors have touched upon epistemological beliefs as further dimensions of scientific literacy and factors influencing students' performance on, for instance, scientific inquiry tasks (Gräber, Nentwig, & Nicolson, 2002; Sandoval, 2005); these beliefs represent another set of covariates against which the authors could evaluate the proposed assessments. One may expect positive relations between these beliefs and science performance so that more sophisticated beliefs about science and scientific knowledge go together with higher performance (e.g., Chen, 2012; Kampa, Neumann, Heitmann, & Kremer, 2016; Mason, Boscolo, Tornatora, & Ronconi, 2013). Motivational, volitional, or even personality-related constructs such as scientific self-concept and self-efficacy, the willingness to engage in scientific problems, the openness to acquire and apply content knowledge, or the perseverance to work on scientific inquiry tasks are further candidates for validation purposes (e.g., Duckworth & Yeager, 2015; Jansen, Scherer, & Schroeders, 2015). Information about the relations to these constructs may provide insights into what determines task performance or response processes and further strengthen the evidence on cognitive and inferential validity (e.g., Goldhammer et al., 2014; Greiff, Niepel, Scherer, & Martin, 2016; Pellegrino et al., 2016).

Implications for the teaching and learning of science – Instructional validity

Kampa and Köller (2016) provide evidence on inferential validity and some indicators of cognitive validity. This perspective can be extended to instructional validity that provides

information about science assessments and has probably the closest connection to teaching and learning. Nonetheless, gathering evidence on instructional validity is by no means as straightforward as examining the internal structure or relations to other constructs. It needs complex study designs and a set of methodological approaches that connect the scale, item, and person characteristics with information about instruction (Naumann, Hochweber, & Klieme, 2016; Polikoff, 2010). Despite these challenges, information on instructional validity is valuable to researchers, teachers, and policy-makers, because it shows how instruction or instructional changes are related to students' performance on science assessments (Pellegrino et al., 2016). Along these lines, I point to potential directions for further research.

I suggest taking an instructional validity perspective that points at the degree to which the proposed assessments are sensitive to the instruction students receive in their science classes. Polikoff (2010) explained that this perspective refers to what is often called "instructional sensitivity" and proposed different approaches to it. The question in any study that attempts to validate an assessment in light of instruction consequently is: To what extent can students' responses and/or the task characteristics be linked to the instructional approaches, or opportunities? If a link can be established, the value of the assessment (i.e., in being sensitive to instruction or instructional changes) and the resultant scale or item scores (i.e., in being indicative of the proficiency in curricular competences or the result of curricular changes) can be clearly recognized. Given that Kampa and Köller (2016) already indicated that the curricular demands with respect to the concept of biological literacy and the assessment contents are well-aligned, there is a great chance that the proposed assessments are instructionally sensitive. Even further, if evidence that the internal structure remains invariant as students move along the grade levels and the curriculum can be obtained, progressions in content knowledge and scientific inquiry skills could be identified (Fortus & Krajcik, 2012; Köller & Parchmann, 2012). This evidence is highly relevant for policy-

makers – because it provides insights into effects of instructional approaches or curricular changes (Duckworth & Yeager, 2015).

Extending the view on validity – Further sources of evidence based on alternative data collection activities

While the sources for crafting a validity argument primarily focused on the performance of items or tests and potential ways to link test scores to instructional aspects, other, more qualitative approaches may enhance the validity argument even further and, at the same time, provide meaningful insights into the functioning of items or a test. In particular, Pellegrino et al. (2016) proposed data collection activities that focus on different participants during the test development process. These activities include:

- *Expert analyses* aimed at (a) reviewing the cognitive demands, ethnic and cultural sensitivity of items or the test; (b) examining the alignment of the test design with instructional uses and needs and the promotion of teacher understanding; (c) assessing the performance of scoring rubrics and/or inferential models,
- *Student cognitive protocol studies* aimed at (a) evaluating which test scores reflect students' thinking processes and proficiencies; (b) reviewing the interaction between test scores and instructional goals and how assessment outcomes support teachers' instructional decisions; (c) examining how students' engagement with assessment activities support analytic models and the corresponding parameters (e.g., relations to cognate or distinct constructs),
- *Teacher studies* aimed at (a) examining the alignment among teachers' interpretation of students outcomes, the design, and intent of the assessment; (b) evaluating teachers' understanding of assessment outcomes and their decisions on instructional use; (c) evaluating teachers' knowledge about and use of the

psychometric properties of items and assessments (e.g., item difficulties, score reliability).

These three data collection activities are by no means exhaustive, and researchers as well as test developers may follow goals other than the ones mentioned here. Still, the main purpose of these activities is to gain a deeper understanding of the consequences related to assessment activities – be it students' thinking processes or teachers' use of test scores. The additional information is valuable to supplement more "quantitative" evidence used for crafting a validity argument. I believe that Kampa's and Köller's (2016) study paves the way for an in-depth evaluation of students' cognitive processes involved in the evaluation of content knowledge and scientific inquiry in biology.

Conclusion

Overall, I agree with Kampa and Köller (2016) insofar that crafting an argument for the validity of science assessments – particularly for those used to develop national proficiency scales – demands evidence on the internal structure and the relations to other constructs. Their paper exemplifies good practice of both obtaining and presenting validity evidence based on solid conceptual ground. In order to craft a validity argument, researchers can make use of further sources of evidence, such as the evaluation of differential item functioning across relevant groups of students, relations to "non-cognitive" constructs, and instructional sensitivity of the measures. These considerations provide science educators with valuable information on the assessment and the construct such that more direct implications for the teaching and learning of science can be derived. My rather general plea is to extend sources of validity evidence of science assessments in order to make a step further in the quest for the "holy grail" of validity.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association (AERA).

Babiar, T. C. (2010). Exploring differential item functioning (DIF) with the Rasch model: a comparison of gender differences on eighth grade science items in the United States and Spain. *J Appl Meas, 12*(2), 144-164.

Bauer, D. J. (2016). A More General Model for Testing Measurement Invariance and Differential Item Functioning. *Psychological Methods*. doi:10.1037/met0000077

Chen, J. A. (2012). Implicit theories, epistemic beliefs, and science motivation: A person-centered approach. *Learning and Individual Differences, 22*(6), 724-735. doi:10.1016/j.lindif.2012.07.013

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237-251. doi:10.3102/0013189X15584327

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*(11), 2417-2425. doi:10.1037/a0037996

Eggert, S., & Bögeholz, S. (2010). Students' use of decision-making strategies with regard to socioscientific issues: An application of the Rasch partial credit model. *Science Education, 94*(2), 230-258. doi:10.1002/sce.20358

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science, 8*(3), 223-241. doi:10.1177/1745691612460685

Fortus, D., & Krajcik, J. (2012). Curriculum Coherence and Learning Progressions. In J. B. Fraser, K. Tobin, & J. C. McRobbie (Eds.), *Second International Handbook of Science Education* (pp. 783-798). Dordrecht: Springer Netherlands.

Goldhammer, F., Naumann, J., Stelter, A., Toacute, th, K., Rodie, . . . Klieme, E. (2014). The Time on Task Effect in Reading and Problem Solving Is Moderated by Task Difficulty and Skill: Insights From a Computer-Based Large-Scale Assessment. *Journal of Educational Psychology, 106*(3), 608-626. doi:10.1037/a0034716

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36-46. doi:10.1016/j.chb.2016.02.095

Gräber, W., Nentwig, P., & Nicolson, P. (2002). Scientific literacy - Von der Theorie zur Praxis [From Theory to Practice]. In W. Gräber, P. Nentwig, T. Koballa, & R. H. Evans (Eds.), *Scientific literacy. Der Beitrag der Naturwissenschaften zur Allgemeinen Bildung* (pp. 135-145). Opladen: Lesle+Budrich.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*(2–3), 57-63. doi:10.1016/j.stueduc.2009.10.002

Jansen, M., Scherer, R., & Schroeders, U. (2015). Students' self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology, 41*, 13-24. doi:10.1016/j.cedpsych.2014.11.002

Kampa, N., & Köller, O. (2016). German National Proficiency Scales in Biology: Internal Structure, Relations to General Cognitive Abilities and Verbal Skills. *Science Education, 100,* 903-922. doi:10.1002/sce.21227

Kampa, N., Neumann, I., Heitmann, P., & Kremer, K. (2016). Epistemological beliefs in science - A person-centered approach to investigate high school students' profiles. *Contemporary Educational Psychology, 46*, 81-93. doi:10.1016/j.cedpsych.2016.04.007

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development, 23*(4), 435-451. doi:10.1016/j.cogdev.2008.09.006

Köller, O., & Parchmann, I. (2012). Competencies: The German notion of learning outcomes *Making it tangible–Learning outcomes in science education* (pp. 165-185). Münster: Waxmann.

Lee, V. E., & Burkam, D. T. (1996). Gender differences in middle grade science achievement: Subject domain, ability level, and course emphasis. *Science Education, 80*(6), 613-650. doi:10.1002/(SICI)1098-237X(199611)80:6<613::AID-SCE1>3.0.CO;2-M

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*: Cambridge University Press.

Liu, X. (2010). *Using and developing measurement instruments in science education*. Charlotte, NC: Information Age Publishing.

Mason, L., Boscolo, P., Tornatora, M. C., & Ronconi, L. (2013). Besides knowledge: a cross-sectional study on the relations between epistemic beliefs, achievement goals, self-beliefs, and achievement in science. *Instructional Science, 41*(1), 49-79. doi:10.1007/s11251-012-9210-0

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Naumann, A., Hochweber, J., & Klieme, E. (2016). A Psychometric Framework for the Evaluation of Instructional Sensitivity. *Educational Assessment*, 0-0. doi:10.1080/10627197.2016.1167591

OECD. (2015). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence*. Paris: OECD Publishing.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educational Psychologist, 51*(1), 59-81. doi:10.1080/00461520.2016.1145550

Polikoff, M. A. (2010). Instructional Sensitivity as a Psychometric Property of Assessments. *Educational Measurement: Issues and Practice, 29*(4), 3-14. doi:10.1111/j.1745-3992.2010.00189.x

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE Life Sciences Education, 15*, 1-9. doi:10.1187/cbe.15-08-0183

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education, 89*(4), 634-656. doi:10.1002/sce.20065

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Mahwah, NJ: Lawrence Erlbaum.

Wang, M.-T., & Degol, J. L. (2016). Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions. *Educational Psychology Review*, 1-22. doi:10.1007/s10648-015-9355-x

Williams, D. C., Ma, Y., Prejean, L., Ford, M. J., & Lai, G. (2007). Acquisition of Physics Content Knowledge and Scientific Inquiry Skills in a Robotics Summer Camp.

Journal of Research on Technology in Education, 40(2), 201-216. doi:10.1080/15391523.2007.10782505

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wind, S. A., & Gale, J. D. (2015). Diagnostic Opportunities Using Rasch Measurement in the Context of a Misconceptions-Based Physical Science Assessment. *Science Education, 99*(4), 721-741. doi:10.1002/sce.21172

Zumbo, B. D. (2006). Validity: Foundational Issues and Statistical Methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26, pp. 45-79). Amsterdam: Elsevier.

Zwick, R. (2012). A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. *ETS Research Report Series, 2012*, i-30. doi:10.1002/j.2333-8504.2012.tb02290.x

Tables

Table 1

Potential Sources of Evidence for Validity Components Based on Studies of Item and Test Performance. Adapted from Pellegrino et al. (2016, p. 68)

Cognitive validity	*Instructional validity*	*Inferential validity*
Extent to which item and test performance support the underlying cognitive processing demands: ▪ Fit between test scores and underlying cognition ▪ Fit between item scores and underlying cognition	Extent to which assessment outcomes support instructional needs: ▪ Formative use of the assessment ▪ Summative monitoring of progress ▪ Connections to external assessments (e.g., examinations)	Extent to which model-based analyses support the intended purpose and use of the assessment: ▪ Scale score and diagnostic reliability ▪ Fit between the model and the data ▪ Dimensionality of the assessment ▪ Differential item functioning for linguistic, ethnic, and other relevant groups ▪ Predictive validity ▪ Alignment with other tests