frontiers in Education

ORIGINAL RESEARCH
published: 08 September 2020
doi: 10.3389/feduc.2020.00156

# Performance Assessment of Critical Thinking: Conceptualization, Design, and Implementation

*Henry I. Braun[1]*, Richard J. Shavelson[2], Olga Zlatkin-Troitschanskaia[3] and Katrina Borowiec[1]*

*[1] Lynch School of Education and Human Development, Boston College, Chestnut Hill, MA, United States, [2] Graduate School of Education, Stanford University, Stanford, CA, United States, [3] Department of Business and Economics Education, Johannes Gutenberg University, Mainz, Germany*

Enhancing students' critical thinking (CT) skills is an essential goal of higher education. This article presents a systematic approach to conceptualizing and measuring CT. CT generally comprises the following mental processes: identifying, evaluating, and analyzing a problem; interpreting information; synthesizing evidence; and reporting a conclusion. We further posit that CT also involves dealing with dilemmas involving ambiguity or conflicts among principles and contradictory information. We argue that performance assessment provides the most realistic—and most credible—approach to measuring CT. From this conceptualization and construct definition, we describe one possible framework for building performance assessments of CT with attention to extended performance tasks within the assessment system. The framework is a product of an ongoing, collaborative effort, the *International Performance Assessment of Learning* (iPAL). The framework comprises four main aspects: (1) The storyline describes a carefully curated version of a complex, real-world situation. (2) The challenge frames the task to be accomplished (3). A portfolio of documents in a range of formats is drawn from multiple sources chosen to have specific characteristics. (4) The scoring rubric comprises a set of scales each linked to a facet of the construct. We discuss a number of use cases, as well as the challenges that arise with the use and valid interpretation of performance assessments. The final section presents elements of the iPAL research program that involve various refinements and extensions of the assessment framework, a number of empirical studies, along with linkages to current work in online reading and information processing.

**Keywords: critical thinking, performance assessment, assessment framework, scoring rubric, evidence-centered design, 21st century skills, higher education**

OPEN ACCESS

***Edited by:***
*Isabel Benítez, University of Granada, Spain*

***Reviewed by:***
*Anders Jönsson, Kristianstad University, Sweden*
*Katrina Roohr, Educational Testing Service, United States*

****Correspondence:***
*Henry I. Braun*
*braunh@bc.edu*

***Specialty section:***
*This article was submitted to Educational Psychology, a section of the journal Frontiers in Education*

***Received:*** *30 May 2020*
***Accepted:*** *04 August 2020*
***Published:*** *08 September 2020*

***Citation:***
*Braun HI, Shavelson RJ, Zlatkin-Troitschanskaia O and Borowiec K (2020) Performance Assessment of Critical Thinking: Conceptualization, Design, and Implementation. Front. Educ. 5:156. doi: 10.3389/feduc.2020.00156*

## INTRODUCTION

In their mission statements, most colleges declare that a principal goal is to develop students' higher-order cognitive skills such as critical thinking (CT) and reasoning (e.g., Shavelson, 2010; Hyytinen et al., 2019). The importance of CT is echoed by business leaders (Association of American Colleges and Universities [AACU], 2018), as well as by college faculty (for curricular analyses in Germany, see e.g., Zlatkin-Troitschanskaia et al., 2018). Indeed, in the 2019 administration of the Faculty Survey of Student Engagement (FSSE), 93% of faculty

reported that they "very much" or "quite a bit" structure their courses to support student development with respect to thinking critically and analytically. In a listing of 21st century skills, CT was the most highly ranked among FSSE respondents (Indiana University, 2019). Nevertheless, there is considerable evidence that many college students do not develop these skills to a satisfactory standard (Arum and Roksa, 2011; Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019). This state of affairs represents a serious challenge to higher education – and to society at large.

In view of the importance of CT, as well as evidence of substantial variation in its development during college, its proper measurement is essential to tracking progress in skill development and to providing useful feedback to both teachers and learners. Feedback can help focus students' attention on key skill areas in need of improvement, and provide insight to teachers on choices of pedagogical strategies and time allocation. Moreover, comparative studies at the program and institutional level can inform higher education leaders and policy makers.

The conceptualization and definition of CT presented here is closely related to models of information processing and online reasoning, the skills that are the focus of this special issue. These two skills are especially germane to the learning environments that college students experience today when much of their academic work is done online. Ideally, students should be capable of more than naïve Internet search, followed by copy-and-paste (e.g., McGrew et al., 2017); rather, for example, they should be able to critically evaluate both sources of evidence and the quality of the evidence itself in light of a given purpose (Leu et al., 2020).

In this paper, we present a systematic approach to conceptualizing CT. From that conceptualization and construct definition, we present one possible framework for building *performance assessments* of CT with particular attention to extended performance tasks within the test environment. The penultimate section discusses some of the challenges that arise with the use and valid interpretation of performance assessment scores. We conclude the paper with a section on future perspectives in an emerging field of research – the iPAL program.

## CONCEPTUAL FOUNDATIONS, DEFINITION AND MEASUREMENT OF CRITICAL THINKING

In this section, we briefly review the concept of CT and its definition. In accordance with the principles of evidence-centered design (ECD; Mislevy et al., 2003), the conceptualization drives the measurement of the construct; that is, implementation of ECD directly links aspects of the assessment framework to specific facets of the construct. We then argue that performance assessments designed in accordance with such an assessment framework provide the most realistic—and most credible—approach to measuring CT. The section concludes with a sketch of an approach to CT measurement grounded in *performance assessment*.

### Concept and Definition of Critical Thinking

Taxonomies of 21st century skills (Pellegrino and Hilton, 2012) abound, and it is neither surprising that CT appears in most taxonomies of learning, nor that there are many different approaches to defining and operationalizing the construct of CT. There is, however, general agreement that CT is a multifaceted construct (Liu et al., 2014). Liu et al. (2014) identified five key facets of CT: (i) evaluating evidence and the use of evidence; (ii) analyzing arguments; (iii) understanding implications and consequences; (iv) developing sound arguments; and (v) understanding causation and explanation.

There is empirical support for these facets from college faculty. A 2016–2017 survey conducted by the Higher Education Research Institute (HERI) at the University of California, Los Angeles found that a substantial majority of faculty respondents "frequently" encouraged students to: (i) evaluate the quality or reliability of the information they receive; (ii) recognize biases that affect their thinking; (iii) analyze multiple sources of information before coming to a conclusion; and (iv) support their opinions with a logical argument (Stolzenberg et al., 2019).

There is general agreement that CT involves the following mental processes: identifying, evaluating, and analyzing a problem; interpreting information; synthesizing evidence; and reporting a conclusion (e.g., Erwin and Sebrell, 2003; Kosslyn and Nelson, 2017; Shavelson et al., 2018). We further suggest that CT includes dealing with dilemmas of ambiguity or conflict among principles and contradictory information (Oser and Biedermann, 2020).

Importantly, Oser and Biedermann (2020) posit that CT can be manifested at three levels. The first level, *Critical Analysis*, is the most complex of the three levels. Critical Analysis requires both knowledge in a specific discipline (conceptual) and procedural analytical (deduction, inclusion, etc.) knowledge. The second level is *Critical Reflection*, which involves more generic skills "... necessary for every responsible member of a society" (p. 90). It is "a basic attitude that must be taken into consideration if (new) information is questioned to be true or false, reliable or not reliable, moral or immoral etc." (p. 90). To engage in Critical Reflection, one needs not only apply analytic reasoning, but also adopt a reflective stance toward the political, social, and other consequences of choosing a course of action. It also involves analyzing the potential motives of various actors involved in the dilemma of interest. The third level, *Critical Alertness*, involves questioning one's own or others' thinking from a skeptical point of view.

Wheeler and Haertel (1993) categorized higher-order skills, such as CT, into two types: (i) when solving problems and making decisions in professional and everyday life, for instance, related to civic affairs and the environment; and (ii) in situations where various mental processes (e.g., comparing, evaluating, and justifying) are developed through formal instruction, usually in a discipline. Hence, in both settings, individuals must confront situations that typically involve a problematic event, contradictory information, and possibly conflicting principles. Indeed, there is an ongoing debate concerning whether CT

should be evaluated using generic or discipline-based assessments (Nagel et al., 2020). Whether CT skills are conceptualized as generic or discipline-specific has implications for how they are assessed and how they are incorporated into the classroom.

In the iPAL project, CT is characterized as a multifaceted construct that comprises conceptualizing, analyzing, drawing inferences or synthesizing information, evaluating claims, and applying the results of these reasoning processes to various purposes (e.g., solve a problem, decide on a course of action, find an answer to a given question or reach a conclusion) (Shavelson et al., 2019). In the course of carrying out a CT task, an individual typically engages in activities such as specifying or clarifying a problem; deciding what information is relevant to the problem; evaluating the trustworthiness of information; avoiding judgmental errors based on "fast thinking"; avoiding biases and stereotypes; recognizing different perspectives and how they can reframe a situation; considering the consequences of alternative courses of actions; and communicating clearly and concisely decisions and actions. The order in which activities are carried out can vary among individuals and the processes can be non-linear and reciprocal.

In this article, we focus on generic CT skills. The importance of these skills derives not only from their utility in academic and professional settings, but also the many situations involving challenging moral and ethical issues – often framed in terms of conflicting principles and/or interests – to which individuals have to apply these skills (Kegan, 1994; Tessier-Lavigne, 2020). Conflicts and dilemmas are ubiquitous in the contexts in which adults find themselves: work, family, civil society. Moreover, to remain viable in the global economic environment – one characterized by increased competition and advances in second generation artificial intelligence (AI) – today's college students will need to continually develop and leverage their CT skills. Ideally, colleges offer a supportive environment in which students can develop and practice effective approaches to reasoning about and acting in learning, professional and everyday situations.

## Measurement of Critical Thinking

Critical thinking is a multifaceted construct that poses many challenges to those who would develop relevant and valid assessments. For those interested in current approaches to the measurement of CT that are not the focus of this paper, consult Zlatkin-Troitschanskaia et al. (2018).

In this paper, we have singled out *performance assessment* as it offers important advantages to measuring CT. Extant tests of CT typically employ response formats such as forced-choice or short-answer, and scenario-based tasks (for an overview, see Liu et al., 2014). They all suffer from moderate to severe construct underrepresentation; that is, they fail to capture important facets of the CT construct such as perspective taking and communication. High fidelity performance tasks are viewed as more authentic in that they provide a problem context and require responses that are more similar to what individuals confront in the real world than what is offered by traditional multiple-choice items (Messick, 1994; Braun, 2019). This greater verisimilitude promises higher levels of construct representation and lower levels of construct-irrelevant variance. Such performance tasks have the capacity to measure facets of CT that are imperfectly assessed, if at all, using traditional assessments (Lane and Stone, 2006; Braun, 2019; Shavelson et al., 2019). However, these assertions must be empirically validated, and the measures should be subjected to psychometric analyses. Evidence of the reliability, validity, and interpretative challenges of performance assessment (PA) are extensively detailed in Davey et al. (2015).

We adopt the following definition of performance assessment:

> A performance assessment (sometimes called a work sample when assessing job performance) ... is an activity or set of activities that requires test takers, either individually or in groups, to generate products or performances in response to a complex, most often real-world task. These products and performances provide observable evidence bearing on test takers' knowledge, skills, and abilities—their competencies—in completing the assessment (Davey et al., 2015, p. 10).

A performance assessment typically includes an extended performance task and short constructed-response and selected-response (i.e., multiple-choice) tasks (for examples, see Zlatkin-Troitschanskaia and Shavelson, 2019). In this paper, we refer to both individual performance- and constructed-response tasks as performance tasks (PT) (For an example, see **Table 1** in section "iPAL Assessment Framework").

# AN APPROACH TO PERFORMANCE ASSESSMENT OF CRITICAL THINKING: THE IPAL PROGRAM

The approach to CT presented here is the result of ongoing work undertaken by the International Performance Assessment of Learning collaborative (iPAL[1]). iPAL is an international consortium of volunteers, primarily from academia, who have come together to address the dearth in higher education of research and practice in measuring CT with performance tasks (Shavelson et al., 2018). In this section, we present iPAL's assessment framework as the basis of measuring CT, with examples along the way.

## iPAL Background

The iPAL assessment framework builds on the Council of Aid to Education's Collegiate Learning Assessment (CLA). The CLA was designed to measure cross-disciplinary, generic competencies, such as CT, analytic reasoning, problem solving, and written communication (Klein et al., 2007; Shavelson, 2010). Ideally, each PA contained an extended PT (e.g., examining a range of evidential materials related to the crash of an aircraft) and two short PT's: one in which students either critique an argument or provide a solution in response to a real-world societal issue.

Motivated by considerations of adequate reliability, in 2012, the CLA was later modified to create the CLA+. The CLA+ includes two subtests: a PT and a 25-item Selected Response

[1]https://www.ipal-rd.com/

**TABLE 1 |** The iPAL assessment framework.

| Aspect | Description | Refugee crisis exemplar |
|---|---|---|
| Storyline | The storyline describes a curated version of a real-world situation. | With regional economic, health, crime and political challenges, there is an increasing demand for migrant entry into the country of Dorado in Central America. The question of whether it is safe to increase immigration (and add to the number of Reception Centers) has come before the country's Homeland Commission. A related question is whether Reception Centers have become local "hotspots" for crime. |
| Challenge | The challenge frames the tasks the respondent must carry out based on the dilemma or problem (potentially including moral or ethical aspects) presented in the *storyline*.<br>The challenge should be sufficiently complex so that its resolution requires the respondent: (i) To apply multiple aspects of reasoning and judgment, and (ii) To consider the trade-offs that occur when adopting one potential solution over another – or deciding among competing principles. | (1a) Enumerate the pros and cons, if any, for accepting more refugees.<br>(1b) Identify the documents and evidence in them to justify the list of pros and cons.<br>(2a) Elaborate and recommend a concrete course of action: stem the flow of refugees at the border, control the flow of refugees (perhaps admitting certain types only like doctors and scientists), or take in a quota decided upon by the inter-governmental agreements.<br>(2b) Identify the documents and evidence in them that lead to the recommendation.<br>(3) Provide a set of recommendations on how the country can address challenges of the poor conditions to which refugees are now exposed, as well as dealing with crime rates in or near Reception Centers.<br>(4) Suggest what additional information, if any, you would like to have to increase your confidence in the recommendation. |
| Documents | The storyline is augmented by a portfolio of documents in a range of formats (e.g., government reports, newspaper articles, web blogs, YouTube videos).<br>Documents are collected or developed purposively to represent different sources of information and multiple perspectives.<br>They vary with respect to the trustworthiness of the information; the relevance of the information; and the extent to which the information provided provokes the respondent to make judgmental errors or show bias. | (1) A letter from the Director of the Valparaiso Metropolitan Reception Center titled "Need for Reception Centers in Crisis Situations."<br>(2) Three tables displaying crime statistics and demographic data provided by the Doradian Bureau of Statistics, presented separately for El Doradians and "foreigners."<br>(3) An interview regarding the integration of migrants with a professor who is an expert on migration.<br>(4) A newspaper article titled "Crimes committed by foreigners are on the rise."<br>(5) An excerpt from a 2016 government report titled "Immigration and security: current status and future predictions."<br>(6) Excerpt from the United Nation's "Universal Declaration of Human Rights."<br>(7) Extract from an OECD Migration Report. |
| Scoring rubric | The scoring rubric comprises six dimensions.<br>The first three dimensions involve comparing, evaluating, and justifying the characteristics of the information provided in the document collection regarding:<br>(1) **Trustworthiness** of the information—dealing primarily with the information source, its context, its (hidden) motivation, and its potential conflicts with other evidence.<br>(2) **Relevance** of the information as it pertains to the problem in the storyline.<br>(3) **Bias** in information due to susceptibility to bias or proneness to use faulty heuristics in judgment and decision-making.<br>The last three dimensions pertain to tacit and explicit response processes:<br>(4) Analysis of different **perspectives** at play, addresses questions about the source of (hidden) motivation, control, expertise, and legitimacy (Mejía et al., 2019).<br>(5) Demonstrating an **openness to the consequences** of prioritizing certain perspectives in the source provided —including any course of action suggested by the materials.<br>(6) Formulating and communicating a **coherent argument** for the position taken, drawing from the five dimensions above. | **Refugee Crisis: Trustworthiness, Relevance, Bias, and Ethical Considerations in Documents**<br>**Document 1**: A letter from the Director of the private reception center (both relevant and irrelevant, *baseline heuristic*).<br>**Document 2**: Doradian Bureau of Statistics – Crime statistics (relevant, *representative and baseline heuristics*).<br>**Document 3**: An interview with a professor who is an expert on immigration (relevant/focuses on the key factors influencing on the success of integration).<br>**Document 4**: Newspaper story (irrelevant, *biased/fake news*).<br>**Document 5**: Government report (relevant).<br>**Document 6**: The United Nations, The Universal Declaration of Human Rights (relevant).<br>**Document 7**: A graph/table from an OECD report with data bearing on increase in refugees and non-refugees and crime (irrelevant, *biased*). |

Question (SRQ) section. The PT presents a document or problem statement and an assignment based on that document which elicits an open-ended response. The CLA+ added the SRQ section (which is not linked substantively to the PT scenario) to increase the number of student responses to obtain more reliable estimates of performance at the student-level than could be achieved with a single PT (Zahner, 2013; Davey et al., 2015).

## iPAL Assessment Framework

### Methodological Foundations

The iPAL framework evolved from the Collegiate Learning Assessment developed by Klein et al. (2007). It was also informed by the results from the AHELO pilot study (Organisation for Economic Co-operation and Development [OECD], 2012, 2013), as well as the KoKoHs research program in Germany

(for an overview see, Zlatkin-Troitschanskaia et al., 2017, 2020). The ongoing refinement of the iPAL framework has been guided in part by the principles of Evidence Centered Design (ECD) (Mislevy et al., 2003; Mislevy and Haertel, 2006; Haertel and Fujii, 2017).

In educational measurement, an assessment framework plays a critical intermediary role between the theoretical formulation of the construct and the development of the assessment instrument containing tasks (or items) intended to elicit evidence with respect to that construct (Mislevy et al., 2003). Builders of the assessment framework draw on the construct theory and operationalize it in a way that provides explicit guidance to PT's developers. Thus, the framework should reflect the relevant facets of the construct, where relevance is determined by substantive theory or an appropriate alternative such as behavioral samples from real-world situations of interest (criterion-sampling; McClelland, 1973), as well as the intended use(s) (for an example, see Shavelson et al., 2019). By following the requirements and guidelines embodied in the framework, instrument developers strengthen the claim of construct validity for the instrument (Messick, 1994).

An assessment framework can be specified at different levels of granularity: an assessment battery ("omnibus" assessment, for an example see below), a single performance task, or a specific component of an assessment (Shavelson, 2010; Davey et al., 2015). In the iPAL program, a performance assessment comprises one or more extended performance tasks and additional selected-response and short constructed-response items. The focus of the framework specified below is on a single PT intended to elicit evidence with respect to some facets of CT, such as the evaluation of the trustworthiness of the documents provided and the capacity to address conflicts of principles.

From the ECD perspective, an assessment is an instrument for generating information to support an evidentiary argument and, therefore, the intended inferences (claims) must guide each stage of the design process. The construct of interest is operationalized through the *Student Model*, which represents the target knowledge, skills, and abilities, as well as the relationships among them. The student model should also make explicit the assumptions regarding student competencies in foundational skills or content knowledge. The *Task Model* specifies the features of the problems or items posed to the respondent, with the goal of eliciting the evidence desired. The assessment framework also describes the collection of *task models* comprising the instrument, with considerations of construct validity, various psychometric characteristics (e.g., reliability) and practical constraints (e.g., testing time and cost). The student model provides grounds for evidence of validity, especially cognitive validity; namely, that the students are thinking critically in responding to the task(s).

In the present context, the target construct (CT) is the competence of individuals to think critically, which entails solving complex, real-world problems, and clearly communicating their conclusions or recommendations for action based on trustworthy, relevant and unbiased information. The situations, drawn from actual events, are challenging and may arise in many possible settings. In contrast to more reductionist approaches to assessment development, the iPAL approach and framework rests on the assumption that properly addressing these situational demands requires the application of a constellation of CT skills appropriate to the particular task presented (e.g., Shavelson, 2010, 2013). For a PT, the assessment framework must also specify the rubric by which the responses will be evaluated. The rubric must be properly linked to the target construct so that the resulting score profile constitutes evidence that is both relevant and interpretable in terms of the student model (for an example, see Zlatkin-Troitschanskaia et al., 2019).

## iPAL Task Framework

The iPAL 'omnibus' framework comprises four main aspects: A *storyline*, a *challenge*, a *document library*, and a *scoring rubric*. **Table 1** displays these aspects, brief descriptions of each, and the corresponding examples drawn from an iPAL performance assessment (Version adapted from original in Hyytinen and Toom, 2019). *Storylines* are drawn from various domains; for example, the worlds of business, public policy, civics, medicine, and family. They often involve moral and/or ethical considerations. Deriving an appropriate storyline from a real-world situation requires careful consideration of which features are to be kept *in toto*, which adapted for purposes of the assessment, and which to be discarded. Framing the *challenge* demands care in wording so that there is minimal ambiguity in what is required of the respondent. The difficulty of the *challenge* depends, in large part, on the nature and extent of the information provided in the *document library*, the amount of scaffolding included, as well as the scope of the required response. The amount of information and the scope of the challenge should be commensurate with the amount of time available. As is evident from the table, the characteristics of the documents in the library are intended to elicit responses related to facets of CT. For example, with regard to bias, the information provided is intended to play to judgmental errors due to fast thinking and/or motivational reasoning. Ideally, the situation should accommodate multiple solutions of varying degrees of merit.

The dimensions of the *scoring rubric* are derived from the *Task Model* and *Student Model* (Mislevy et al., 2003) and signal which features are to be extracted from the response and indicate how they are to be evaluated. There should be a direct link between the evaluation of the evidence and the claims that are made with respect to the key features of the *task model* and *student model*. More specifically, the *task model* specifies the various manipulations embodied in the PA and so informs scoring, while the *student model* specifies the capacities students employ in more or less effectively responding to the tasks. The score scales for each of the five facets of CT (see section "Concept and Definition of Critical Thinking") can be specified using appropriate behavioral anchors (for examples, see Zlatkin-Troitschanskaia and Shavelson, 2019). Of particular importance is the evaluation of the response with respect to the last dimension of the scoring rubric; namely, the overall coherence and persuasiveness of the argument, building on the explicit or implicit characteristics related to the first five dimensions. The scoring process must be monitored carefully to

ensure that (trained) raters are judging each response based on the same types of features and evaluation criteria (Braun, 2019) as indicated by interrater agreement coefficients.

The scoring rubric of the iPAL omnibus framework can be modified for specific tasks (Lane and Stone, 2006). This generic rubric helps ensure consistency across rubrics for different storylines. For example, Zlatkin-Troitschanskaia et al. (2019, p. 473) used the following scoring scheme:

> Based on our construct definition of CT and its four dimensions: (D1-Info) recognizing and evaluating information, (D2-Decision) recognizing and evaluating arguments and making decisions, (D3-Conseq) recognizing and evaluating the consequences of decisions, and (D4-Writing), we developed a corresponding analytic dimensional scoring . . . The students' performance is evaluated along the four dimensions, which in turn are subdivided into a total of 23 indicators as (sub)categories of CT . . . For each dimension, we sought detailed evidence in students' responses for the indicators and scored them on a six-point Likert-type scale. In order to reduce judgment distortions, an elaborate procedure of 'behaviorally anchored rating scales' (Smith and Kendall, 1963) was applied by assigning concrete behavioral expectations to certain scale points (Bernardin et al., 1976). To this end, we defined the scale levels by short descriptions of typical behavior and anchored them with concrete examples. . . . We trained four raters in 1 day using a specially developed training course to evaluate students' performance along the 23 indicators clustered into four dimensions (for a description of the rater training, see Klotzer, 2018).

Shavelson et al. (2019) examined the interrater agreement of the scoring scheme developed by Zlatkin-Troitschanskaia et al. (2019) and "found that with 23 items and 2 raters the generalizability ("reliability") coefficient for total scores to be 0.74 (with 4 raters, 0.84)" (Shavelson et al., 2019, p. 15). In the study by Zlatkin-Troitschanskaia et al. (2019, p. 478) three score profiles were identified (low-, middle-, and high-performer) for students. Proper interpretation of such profiles requires care. For example, there may be multiple possible explanations for low scores such as poor CT skills, a lack of a disposition to engage with the challenge, or the two attributes jointly. These alternative explanations for student performance can potentially pose a threat to the evidentiary argument. In this case, auxiliary information may be available to aid in resolving the ambiguity. For example, student responses to selected- and short-constructed-response items in the PA can provide relevant information about the levels of the different skills possessed by the student. When sufficient data are available, the scores can be modeled statistically and/or qualitatively in such a way as to bring them to bear on the technical quality or interpretability of the claims of the assessment: reliability, validity, and utility evidence (Davey et al., 2015; Zlatkin-Troitschanskaia et al., 2019). These kinds of concerns are less critical when PT's are used in classroom settings. The instructor can draw on other sources of evidence, including direct discussion with the student.

## Use of iPAL Performance Assessments in Educational Practice: Evidence From Preliminary Validation Studies

The assessment framework described here supports the development of a PT in a general setting. Many modifications are possible and, indeed, desirable. If the PT is to be more deeply embedded in a certain discipline (e.g., economics, law, or medicine), for example, then the framework must specify characteristics of the narrative and the complementary documents as to the breadth and depth of disciplinary knowledge that is represented.

At present, preliminary field trials employing the omnibus framework (i.e., a full set of documents) indicated that 60 min was generally an inadequate amount of time for students to engage with the full set of complementary documents and to craft a complete response to the challenge (for an example, see Shavelson et al., 2019). Accordingly, it would be helpful to develop modified frameworks for PT's that require substantially less time. For an example, see a short performance assessment of civic online reasoning, requiring response times from 10 to 50 min (Wineburg et al., 2016). Such assessment frameworks could be derived from the omnibus framework by focusing on a reduced number of facets of CT, and specifying the characteristics of the complementary documents to be included – or, perhaps, choices among sets of documents. In principle, one could build a 'family' of PT's, each using the same (or nearly the same) storyline and a subset of the full collection of complementary documents.

Paul and Elder (2007) argue that the goal of CT assessments should be to provide faculty with important information about how well their instruction supports the development of students' CT. In that spirit, the full family of PT's could represent all facets of the construct while affording instructors and students more specific insights on strengths and weaknesses with respect to particular facets of CT. Moreover, the framework should be expanded to include the design of a set of short answer and/or multiple choice items to accompany the PT. Ideally, these additional items would be based on the same narrative as the PT to collect more nuanced information on students' precursor skills such as reading comprehension, while enhancing the overall reliability of the assessment. Areas where students are underprepared could be addressed before, or even in parallel with the development of the focal CT skills. The parallel approach follows the co-requisite model of developmental education. In other settings (e.g., for summative assessment), these complementary items would be administered after the PT to augment the evidence in relation to the various claims. The full PT taking 90 min or more could serve as a capstone assessment.

As we transition from simply delivering paper-based assessments by computer to taking full advantage of the affordances of a digital platform, we should learn from the hard-won lessons of the past so that we can make swifter progress with fewer missteps. In that regard, we must take validity as the touchstone – assessment design, development and deployment must all be tightly linked to the operational definition of the CT construct. Considerations of reliability and practicality come into play with various use cases that highlight different purposes for the assessment (for future perspectives, see next section).

The iPAL assessment framework represents a feasible compromise between commercial, standardized assessments of CT (e.g., Liu et al., 2014), on the one hand, and, on the other, freedom for individual faculty to develop assessment tasks according to idiosyncratic models. It imposes a degree of standardization on *both* task development and scoring, while still allowing some flexibility for faculty to tailor the assessment to meet their unique needs. In so doing, it addresses a key weakness of the AAC&U's VALUE initiative[2] (retrieved 5/7/2020) that has achieved wide acceptance among United States colleges.

The VALUE initiative has produced generic scoring rubrics for 15 domains including CT, problem-solving and written communication. A rubric for a particular skill domain (e.g., critical thinking) has five to six dimensions with four ordered performance levels for each dimension (1 = lowest, 4 = highest). The performance levels are accompanied by language that is intended to clearly differentiate among levels.[3] Faculty are asked to submit student work products from a senior level course that is intended to yield evidence with respect to student learning outcomes in a particular domain and that, they believe, can elicit performances at the highest level. The collection of work products is then graded by faculty from other institutions who have been trained to apply the rubrics.

A principal difficulty is that there is neither a common framework to guide the design of the challenge, nor any control on task complexity and difficulty. Consequently, there is substantial heterogeneity in the quality and evidential value of the submitted responses. This also causes difficulties with task scoring and inter-rater reliability. Shavelson et al. (2009) discuss some of the problems arising with non-standardized collections of student work.

In this context, one advantage of the iPAL framework is that it can provide valuable guidance and an explicit structure for faculty in developing performance tasks for both instruction and formative assessment. When faculty design assessments, their focus is typically on content coverage rather than other potentially important characteristics, such as the degree of construct representation and the adequacy of their scoring procedures (Braun, 2019).

# CONCLUDING REFLECTIONS

## Challenges to Interpretation and Implementation

Performance tasks such as those generated by iPAL are attractive instruments for assessing CT skills (e.g., Shavelson, 2010; Shavelson et al., 2019). The attraction mainly rests on the assumption that elaborated PT's are more authentic (direct) and more completely capture facets of the target construct (i.e., possess greater construct representation) than the widely used selected-response tests. However, as Messick (1994) noted authenticity is a "promissory note" that must be redeemed with empirical research. In practice, there are trade-offs among authenticity, construct validity, and psychometric quality such as reliability (Davey et al., 2015).

One reason for Messick (1994) caution is that authenticity does not guarantee construct validity. The latter must be established by drawing on multiple sources of evidence (American Educational Research Association et al., 2014). Following the ECD principles in designing and developing the PT, as well as the associated scoring rubrics, constitutes an important type of evidence. Further, as Leighton (2019) argues, response process data ("cognitive validity") is needed to validate claims regarding the cognitive complexity of PT's. Relevant data can be obtained through cognitive laboratory studies involving methods such as think aloud protocols or eye-tracking. Although time-consuming and expensive, such studies can yield not only evidence of validity, but also valuable information to guide refinements of the PT.

Going forward, iPAL PT's must be subjected to validation studies as recommended in the *Standards for Psychological and Educational Testing* by American Educational Research Association et al. (2014). With a particular focus on the criterion "relationships to other variables," a framework should include assumptions about the theoretically expected relationships among the indicators assessed by the PT, as well as the indicators' relationships to external variables such as intelligence or prior (task-relevant) knowledge.

Complementing the necessity of evaluating construct validity, there is the need to consider potential sources of construct-irrelevant variance (CIV). One pertains to student motivation, which is typically greater when the stakes are higher. If students are not motivated, then their performance is likely to be impacted by factors unrelated to their (construct-relevant) ability (Lane and Stone, 2006; Braun et al., 2011; Shavelson, 2013). Differential motivation across groups can also bias comparisons. Student motivation might be enhanced if the PT is administered in the context of a course with the promise of generating useful feedback on students' skill profiles.

Construct-irrelevant variance can also occur when students are not equally prepared for the format of the PT or fully appreciate the response requirements. This source of CIV could be alleviated by providing students with practice PT's. Finally, the use of novel forms of documentation, such as those from the Internet, can potentially introduce CIV due to differential familiarity with forms of representation or contents. Interestingly, this suggests that there may be a conflict between enhancing construct representation and reducing CIV.

Another potential source of CIV is related to response evaluation. Even with training, human raters can vary in accuracy and usage of the full score range. In addition, raters may attend to features of responses that are unrelated to the target construct, such as the length of the students' responses or the frequency of grammatical errors (Lane and Stone, 2006). Some of these sources of variance could be addressed in an online environment, where word processing software could alert students to potential grammatical and spelling errors before they submit their final work product.

[2]https://www.aacu.org/value

[3]When test results are reported by means of substantively defined categories, the scoring is termed "criterion-referenced". This is, in contrast to results, reported as percentiles; such scoring is termed "norm-referenced".

Performance tasks generally take longer to administer and are more costly than traditional assessments, making it more difficult to reliably measure student performance (Messick, 1994; Davey et al., 2015). Indeed, it is well known that more than one performance task is needed to obtain high reliability (Shavelson, 2013). This is due to both student-task interactions and variability in scoring. Sources of student-task interactions are differential familiarity with the topic (Hyytinen and Toom, 2019) and differential motivation to engage with the task. The level of reliability required, however, depends on the context of use. For use in formative assessment as part of an instructional program, reliability can be lower than use for summative purposes. In the former case, other types of evidence are generally available to support interpretation and guide pedagogical decisions. Further studies are needed to obtain estimates of reliability in typical instructional settings.

With sufficient data, more sophisticated psychometric analyses become possible. One challenge is that the assumption of unidimensionality required for many psychometric models might be untenable for performance tasks (Davey et al., 2015). Davey et al. (2015) provide the example of a mathematics assessment that requires students to demonstrate not only their mathematics skills but also their written communication skills. Although the iPAL framework does not explicitly address students' reading comprehension and organization skills, students will likely need to call on these abilities to accomplish the task. Moreover, as the operational definition of CT makes evident, the student must not only deploy several skills in responding to the challenge of the PT, but also carry out component tasks in sequence. The former requirement strongly indicates the need for a multi-dimensional IRT model, while the latter suggests that the usual assumption of local item independence may well be problematic (Lane and Stone, 2006). At the same time, the analytic scoring rubric should facilitate the use of latent class analysis to partition data from large groups into meaningful categories (Zlatkin-Troitschanskaia et al., 2019).

## Future Perspectives

Although the iPAL consortium has made substantial progress in the assessment of CT, much remains to be done. Further refinement of existing PT's and their adaptation to different languages and cultures must continue. To this point, there are a number of examples: The refugee crisis PT (cited in **Table 1**) was translated and adapted from Finnish to US English and then to Colombian Spanish. A PT concerning kidney transplants was translated and adapted from German to US English. Finally, two PT's based on 'legacy admissions' to US colleges were translated and adapted to Colombian Spanish.

With respect to data collection, there is a need for sufficient data to support psychometric analysis of student responses, especially the relationships among the different components of the scoring rubric, as this would inform both task development and response evaluation (Zlatkin-Troitschanskaia et al., 2019). In addition, more intensive study of response processes through cognitive laboratories and the like are needed to strengthen the evidential argument for construct validity (Leighton, 2019). We are currently conducting empirical studies, collecting data on both iPAL PT's and other measures of CT. These studies will provide evidence of convergent and discriminant validity.

At the same time, efforts should be directed at further development to support different ways CT PT's might be used—i.e., use cases—especially those that call for formative use of PT's. Incorporating formative assessment into courses can plausibly be expected to improve students' competency acquisition (Zlatkin-Troitschanskaia et al., 2017). With suitable choices of storylines, appropriate combinations of (modified) PT's, supplemented by short-answer and multiple-choice items, could be interwoven into ordinary classroom activities. The supplementary items may be completely separate from the PT's (as is the case with the CLA+), loosely coupled with the PT's (as in drawing on the same storyline), or tightly linked to the PT's (as in requiring elaboration of certain components of the response to the PT).

As an alternative to such integration, stand-alone modules could be embedded in courses to yield evidence of students' generic CT skills. Core curriculum courses or general education courses offer ideal settings for embedding performance assessments. If these assessments were administered to a representative sample of students in each cohort over their years in college, the results would yield important information on the development of CT skills at a population level. For another example, these PA's could be used to assess the competence profiles of students entering Bachelor's or graduate-level programs as a basis for more targeted instructional support.

Thus, in considering different use cases for the assessment of CT, it is evident that several modifications of the iPAL omnibus assessment framework are needed. As noted earlier, assessments built according to this framework are demanding with respect to the extensive preliminary work required by a task and the time required to properly complete it. Thus, it would be helpful to have modified versions of the framework, focusing on one or two facets of the CT construct and calling for a smaller number of supplementary documents. The challenge to the student should be suitably reduced.

Some members of the iPAL collaborative have developed PT's that are embedded in disciplines such as engineering, law and education (Crump et al., 2019; for teacher education examples, see Jeschke et al., 2019). These are proving to be of great interest to various stakeholders and further development is likely. Consequently, it is essential that an appropriate assessment framework be established and implemented. It is both a conceptual and an empirical question as to whether a single framework can guide development in different domains.

## Performance Assessment in Online Learning Environment

Over the last 15 years, increasing amounts of time in both college and work are spent using computers and other electronic devices. This has led to formulation of models for the *new literacies* that attempt to capture some key characteristics of these activities. A prominent example is a model proposed by Leu et al. (2020). The model frames online reading as a process of

problem-based inquiry that calls on five practices to occur during online research and comprehension:

1. Reading to identify important questions,
2. Reading to locate information,
3. Reading to critically evaluate information,
4. Reading to synthesize online information, and
5. Reading and writing to communicate online information.

The parallels with the iPAL definition of CT are evident and suggest there may be benefits to closer links between these two lines of research. For example, a report by Leu et al. (2014) describes empirical studies comparing assessments of online reading using either open-ended or multiple-choice response formats.

The iPAL consortium has begun to take advantage of the affordances of the online environment (for examples, see Schmidt et al. and Nagel et al. in this special issue). Most obviously, Supplementary Materials can now include archival photographs, audio recordings, or videos. Additional tasks might include the online search for relevant documents, though this would add considerably to the time demands. This online search could occur within a simulated Internet environment, as is the case for the IEA's ePIRLS assessment (Mullis et al., 2017).

The prospect of having access to a wealth of materials that can add to task authenticity is exciting. Yet it can also add ambiguity and information overload. Increased authenticity, then, should be weighed against validity concerns and the time required to absorb the content in these materials. Modifications of the design framework and extensive empirical testing will be required to decide on appropriate trade-offs. A related possibility is to employ some of these materials in short-answer (or even selected-response) items that supplement the main PT. Response formats could include highlighting text or using a drag-and-drop menu to construct a response. Students' responses could be automatically scored, thereby containing costs. With automated scoring, feedback to students and faculty, including suggestions for next steps in strengthening CT skills, could also be provided without adding to faculty workload. Therefore, taking advantage of the online environment to incorporate new types of supplementary documents should be a high priority and, perhaps, to introduce new response formats as well. Finally, further investigation of the overlap between this formulation of CT and the characterization of online reading promulgated by Leu et al. (2020) is a promising direction to pursue.

## DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

## AUTHOR CONTRIBUTIONS

HB wrote the article. RS, OZ-T, and KB were involved in the preparation and revision of the article and co-wrote the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was funded in part by the Spencer Foundation (Grant No. #201700123).

## ACKNOWLEDGMENTS

We would like to thank all the researchers who have participated in the iPAL program.

## REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, D.C: American Educational Research Association.

Arum, R., and Roksa, J. (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago, IL: University of Chicago Press.

Association of American Colleges and Universities (n.d.). *VALUE: What is value?*. Available online at:: https://www.aacu.org/value (accessed May 7, 2020).

Association of American Colleges and Universities [AACU] (2018). *Fulfilling the American Dream: Liberal Education and the Future of Work*. Available online at:: https://www.aacu.org/research/2018-future-of-work (accessed May 1, 2020).

Braun, H. (2019). Performance assessment and standardization in higher education: a problematic conjunction? *Br. J. Educ. Psychol.* 89, 429–440. doi: 10.1111/bjep.12274

Braun, H. I., Kirsch, I., and Yamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teach. Coll. Rec.* 113, 2309–2344.

Crump, N., Sepulveda, C., Fajardo, A., and Aguilera, A. (2019). Systematization of performance tests in critical thinking: an interdisciplinary construction experience. *Rev. Estud. Educ.* 2, 17–47.

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., and Wise, L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment*. Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service.

Erwin, T. D., and Sebrell, K. W. (2003). Assessment of critical thinking: ETS's tasks in critical thinking. *J. Gen. Educ.* 52, 50–70. doi: 10.1353/jge.2003.0019

Haertel, G. D., and Fujii, R. (2017). "Evidence-centered design and postsecondary assessment," in *Handbook on Measurement, Assessment, and Evaluation in Higher Education*, 2nd Edn, eds C. Secolsky and D. B. Denison (Abingdon: Routledge), 313–339. doi: 10.4324/9781315709307-26

Hyytinen, H., and Toom, A. (2019). Developing a performance assessment task in the Finnish higher education context: conceptual and empirical insights. *Br. J. Educ. Psychol.* 89, 551–563. doi: 10.1111/bjep.12283

Hyytinen, H., Toom, A., and Shavelson, R. J. (2019). "Enhancing scientific thinking through the development of critical thinking in higher education," in *Redefining Scientific Thinking for Higher Education: Higher-Order Thinking, Evidence-Based Reasoning and Research Skills*, eds M. Murtonen and K. Balloo (London: Palgrave MacMillan).

Indiana University (2019). *FSSE 2019 Frequencies: FSSE 2019 Aggregate*. Available online at:: http://fsse.indiana.edu/pdf/FSSE_IR_2019/summary_tables/FSSE19_Frequencies_(FSSE_2019).pdf (accessed May 1, 2020).

Jeschke, C., Kuhn, C., Lindmeier, A., Zlatkin-Troitschanskaia, O., Saas, H., and Heinze, A. (2019). Performance assessment to investigate the domain specificity of instructional skills among pre-service and in-service teachers of mathematics and economics. *Br. J. Educ. Psychol.* 89, 538–550. doi: 10.1111/bjep.12277

Kegan, R. (1994). *In Over Our Heads: The Mental Demands of Modern Life*. Cambridge, MA: Harvard University Press.

Klein, S., Benjamin, R., Shavelson, R., and Bolus, R. (2007). The collegiate learning assessment: facts and fantasies. *Eval. Rev.* 31, 415–439. doi: 10.1177/0193841x07303318

Kosslyn, S. M., and Nelson, B. (2017). *Building the Intentional University: Minerva and the Future of Higher Education*. Cambridge, MAL: The MIT Press.

Lane, S., and Stone, C. A. (2006). "Performance assessment," in *Educational Measurement*, 4th Edn, ed. R. L. Brennan (Lanham, MA: Rowman & Littlefield Publishers), 387–432.

Leighton, J. P. (2019). The risk–return trade-off: performance assessments and cognitive validation of inferences. *Br. J. Educ. Psychol.* 89, 441–455. doi: 10.1111/bjep.12271

Leu, D. J., Kiili, C., Forzani, E., Zawilinski, L., McVerry, J. G., and O'Byrne, W. I. (2020). "The new literacies of online research and comprehension," in *The Concise Encyclopedia of Applied Linguistics*, ed. C. A. Chapelle (Oxford: Wiley-Blackwell), 844–852.

Leu, D. J., Kulikowich, J. M., Kennedy, C., and Maykel, C. (2014). "The ORCA Project: designing technology-based assessments for online research," in *Paper Presented at the American Educational Research Annual Meeting*, Philadelphia, PA.

Liu, O. L., Frankel, L., and Roohr, K. C. (2014). Assessing critical thinking in higher education: current state and directions for next-generation assessments. *ETS Res. Rep. Ser.* 1, 1–23. doi: 10.1002/ets2.12009

McClelland, D. C. (1973). Testing for competence rather than for "intelligence.". *Am. Psychol.* 28, 1–14. doi: 10.1037/h0034092

McGrew, S., Ortega, T., Breakstone, J., and Wineburg, S. (2017). The challenge that's bigger than fake news: civic reasoning in a social media environment. *Am. Educ.* 4, 4-9, 39.

Mejía, A., Mariño, J. P., and Molina, A. (2019). Incorporating perspective analysis into critical thinking performance assessments. *Br. J. Educ. Psychol.* 89, 456–467. doi: 10.1111/bjep.12297

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educ. Res.* 23, 13–23. doi: 10.3102/0013189x023002013

Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Res. Rep. Ser.* 2003, i–29. doi: 10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., and Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educ. Meas. Issues Pract.* 25, 6–20. doi: 10.1111/j.1745-3992.2006.00075.x

Mullis, I. V. S., Martin, M. O., Foy, P., and Hooper, M. (2017). *ePIRLS 2016 International Results in Online Informational Reading*. Available online at:: http://timssandpirls.bc.edu/pirls2016/international-results/ (accessed May 1, 2020).

Nagel, M.-T., Zlatkin-Troitschanskaia, O., Schmidt, S., and Beck, K. (2020). "Performance assessment of generic and domain-specific skills in higher education economics," in *Student Learning in German Higher Education*, eds O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, and C. Lautenbach (Berlin: Springer), 281–299. doi: 10.1007/978-3-658-27886-1_14

Organisation for Economic Co-operation and Development [OECD] (2012). *AHELO: Feasibility Study Report*, Vol. 1. Paris: OECD. Design and implementation.

Organisation for Economic Co-operation and Development [OECD] (2013). *AHELO: Feasibility Study Report*, Vol. 2. Paris: OECD. Data analysis and national experiences.

Oser, F. K., and Biedermann, H. (2020). "A three-level model for critical thinking: critical alertness, critical reflection, and critical analysis," in *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*, ed. O. Zlatkin-Troitschanskaia (Cham: Springer), 89–106. doi: 10.1007/978-3-030-26578-6_7

Paul, R., and Elder, L. (2007). Consequential validity: using assessment to drive instruction. *Found. Crit. Think.* 29, 31–40.

Pellegrino, J. W., and Hilton, M. L. (eds) (2012). *Education for life and work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington DC: National Academies Press.

Shavelson, R. (2010). *Measuring College Learning Responsibly: Accountability in a New Era*. Redwood City, CA: Stanford University Press.

Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educ. Psychol.* 48, 73–86. doi: 10.1080/00461520.2013.779483

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. P. (2019). Assessment of university students' critical thinking: next generation performance assessment. *Int. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309

Shavelson, R. J., Zlatkin-Troitschanskaia, O., and Marino, J. P. (2018). "International performance assessment of learning in higher education (iPAL): research and development," in *Assessment of Learning Outcomes in Higher Education: Cross-National Comparisons and Perspectives*, eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, and C. Kuhn (Berlin: Springer), 193–214. doi: 10.1007/978-3-319-74338-7_10

Shavelson, R. J., Klein, S., and Benjamin, R. (2009). The limitations of portfolios. *Inside Higher Educ.* Available online at: https://www.insidehighered.com/views/2009/10/16/limitations-portfolios

Stolzenberg, E. B., Eagan, M. K., Zimmerman, H. B., Berdan Lozano, J., Cesar-Davis, N. M., Aragon, M. C., et al. (2019). *Undergraduate Teaching Faculty: The HERI Faculty Survey 2016–2017*. Los Angeles, CA: UCLA.

Tessier-Lavigne, M. (2020). *Putting Ethics at the Heart of Innovation*. Stanford, CA: Stanford Magazine.

Wheeler, P., and Haertel, G. D. (1993). *Resource Handbook on Performance Assessment and Measurement: A Tool for Students, Practitioners, and Policymakers*. Palm Coast, FL: Owl Press.

Wineburg, S., McGrew, S., Breakstone, J., and Ortega, T. (2016). *Evaluating Information: The Cornerstone of Civic Online Reasoning. Executive Summary*. Stanford, CA: Stanford History Education Group.

Zahner, D. (2013). *Reliability and Validity–CLA+. Council for Aid to Education*. Available online at:: https://pdfs.semanticscholar.org/91ae/8edfac44bce3bed37d8c9091da01d6db3776.pdf.

Zlatkin-Troitschanskaia, O., and Shavelson, R. J. (2019). Performance assessment of student learning in higher education [Special issue]. *Br. J. Educ. Psychol.* 89, i–iv, 413–563.

Zlatkin-Troitschanskaia, O., Pant, H. A., Lautenbach, C., Molerov, D., Toepper, M., and Brückner, S. (2017). *Modeling and Measuring Competencies in Higher Education: Approaches to Challenges in Higher Education Policy and Practice*. Berlin: Springer VS.

Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., and Lautenbach, C. (eds) (2020). *Student Learning in German Higher Education: Innovative Measurement Approaches and Research Results*. Wiesbaden: Springer.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., and Pant, H. A. (2018). "Assessment of learning outcomes in higher education: international comparisons and perspectives," in *Handbook on Measurement, Assessment, and Evaluation in Higher Education*, 2nd Edn, eds C. Secolsky and D. B. Denison (Abingdon: Routledge), 686–697.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., and Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *Br. J. Educ. Psychol.* 89, 468–484. doi: 10.1111/bjep.12286

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.