DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

PROMOTING LIBERAL EDUCATION THROUGH THE LONGITUDINAL STUDY OF CRITICAL THINKING: A RATIONALE AND PLAN

PROMOVER LA EDUCACIÓN LIBERAL A TRAVÉS DEL ESTUDIO LONGITUDINAL DEL PENSAMIENTO CRÍTICO: FUNDAMENTOS Y PROPUESTA

Henry Braun

Katrina Borowiec

ABSTRACT

In recent years, U.S. higher education has received much criticism for inadequately preparing students for the 'real world.' There is substantial empirical evidence (e.g., Arum & Roksa, 2011) that many students graduate with limited proficiency in key 21st century skills such as critical thinking (CT). Despite its importance in achieving personal advancement and professional success, there has been surprisingly little rigorous research on the development of CT during the undergraduate years. We believe this is a missed opportunity for higher education to better understand the various trajectories of CT development and to generate credible evidence to inform policies, programs, and practices – while also strengthening its standing among its various stakeholders. We argue that, despite a number of challenges, it is feasible to design and implement a comprehensive, longitudinal study of the development of CT (and related constructs). Although CT is an important learning objective at all colleges, it has special resonance for schools and programs that aim to provide a liberal education. For various reasons we propose that the initial study be conducted within a particular subset of such institutions; namely, a sample of American Jesuit colleges and universities. We discuss a number of technical issues germane to such a study, as well as the advantages and disadvantages of our preferred choice of an institutional sample.

Key words: 21st century skills; critical thinking; learning outcomes assessment; liberal education; Jesuit higher education

RESUMEN

Recientemente, la educación superior en Estados Unidos ha recibido bastantes críticas por no preparar de manera adecuada a los estudiantes para "el mundo real". Existe evidencia empírica sustancial (p.ej., Arum & Roksa, 2011) de que muchos estudiantes se gradúan con un nivel limitado de dominio de destrezas del siglo XXI claves, como el Pensamiento Crítico (PC). A pesar de su importancia para el logro del progreso personal y del éxito profesional, sorprendentemente, hay muy poca investigación rigurosa sobre el desarrollo del PC durante la etapa de pregrado. Creemos que esta es una oportunidad perdida en el campo de la educación superior para comprender las variadas trayectorias de desarrollo del PC y para generar evidencia creíble que informe políticas, programas y prácticas profesionales al tiempo que se fortalece el estatus del PC entre los varios interesados en este tema. Sostenemos que, a pesar de los múltiples retos involucrados, es factible diseñar e implementar un estudio longitudinal amplio sobre el desarrollo del PC (y de otros constructos relacionados). Aunque el PC es un objetivo de aprendizaje importante en todas las

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

instituciones de educación superior, tiene una resonancia especial para las instituciones y programas que apuntan a brindar educación liberal. Proponemos que, por varias razones, el estudio inicial debe conducirse en un subconjunto particular de instituciones; a saber, una muestra de instituciones de educación superior Jesuita en los Estados Unidos. Discutimos varios asuntos técnicos relevantes a tal estudio, así como las ventajas y desventajas asociadas a nuestra opción preferida de muestra de instituciones.

Palabras clave: Destrezas del siglo XXI; pensamiento crítico; medición de resultados del aprendizaje; educación liberal; educación superior Jesuita

Date of receipt: 15 march 2021.
Date of acceptance: 22 june 2021.

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

INTRODUCTION

Over the last two decades, U.S. colleges have garnered sustained criticism that they are saddling students with substantial debt while not preparing them for the 'real world.' Although some criticisms are politically motivated and more focused on the supposed superiority of vocationally oriented programs, others are grounded in both anecdotal and empirical evidence that many students are not developing such skills as critical thinking (CT) that are an important learning outcome (Arum & Roksa, 2011; Bok, 2006). A federally sponsored commission decried the fact that colleges face minimal accountability for their students achieving (or not) the learning goals established by the colleges themselves (Spellings, 2006). Other critics deplore the fact that most colleges – even those that characterize themselves as providing a liberal education – have largely abandoned those aspects of their original missions related to supporting their students' holistic growth and development (Delbanco, 2012; Kronman, 2008).

In this climate, one sector of higher education – liberal arts colleges – have attracted particular attention and criticism. Their supporters and, more generally, defenders of liberal education, respond that the learning objectives of a liberal education (e.g., analytic skills, flexibility, openness, and a readiness to engage in lifelong learning) are precisely the skills and dispositions necessary for career success in the 21st century. The American Association of Colleges & Universities (AAC&U) offers many anecdotes to that effect, as well as testimonials from business leaders (AAC&U, 2002). In this climate, CT constitutes a key learning outcome – both in its own right and in its essential role in related, long term outcomes such as workplace success and engaged citizenship (AAC&U, 2020). Indeed, the AAC&U (2020) asserts that a liberally educated citizenry is essential to maintaining a democratic polity and highlights the importance of CT. More generally, whether they are explicit or not, nearly all colleges consider the development of CT skills an important outcome. Results from the 2019 Faculty Survey of Student Engagement indicate that 93% of faculty "quite a bit" or "very much" structure their courses to support the development of CT among their students (Indiana University, 2019). Indeed, many colleges highlight CT as a key outcome (see Figures 1-3).

Figure 1: Grinnell College Spotlight

> Grinnell College, a top liberal arts college, explains that its mission is "to graduate individuals who can think clearly, who can speak and write persuasively and even eloquently, who can *evaluate critically* both their own and others' ideas, who can acquire new knowledge, and who are prepared in life and work to use their knowledge and their abilities to serve the common good" (Grinnell, n.d. para. 1; emphasis added).

Source: Grinnell College (n.d.)

Figure 2: Amsterdam University College Spotlight

> In a report outlining its strategic vision and goals, Amsterdam University College in the Netherlands described one of its priorities as "Affirming our commitment to a curriculum built round the research cycle and the key learning objectives of LAS [Liberal Arts and Sciences], such as systematic literature review, formulating coherent research questions and writing, experimentation, *critical analysis*, data processing and interpretation, interdisciplinary integration, self-reflection and *critical analysis*, and dissemination for a wider audience" (Pratt, 2016, p. 15 ; emphasis added).

Source: Pratt (2016)

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

Figure 3: Boston College Spotlight

> Boston College is a noteworthy exemplar of periodic recommitment to the ideals of a liberal education in the context of a faith-based institution. In a seminal document, (Boston College, 2007), Boston College envisions development as proceeding along three inter-related strands: Intellectual, social, and spiritual. Most importantly, it asserts that the progressive integration of these strands into a coherent whole is the mark of a mature individual.

Source: Boston College (2007)

However, despite the centrality of CT, there has been relatively little rigorous research tracking the development of CT through the undergraduate years. There are certainly many methodological and logistical challenges to conducting the necessary longitudinal studies. Nonetheless, we believe that higher education, and programs in liberal education in particular, can ill-afford the status quo. We propose, therefore, that higher education affirmatively support the initiation, design and implementation of a comprehensive study to document the trajectories of CT development.

Admittedly, CT is a complex, multi-faceted construct (Braun et al., 2020; Liu et al., 2014; Stassen et al., 2011), which makes it challenging—but certainly not impossible—to measure, as we discuss below. In fact, there has been considerable research on developing operational definitions of CT and on crafting assessments to measure different facets of the construct (Anghel et al., in press; Braun et al., 2020; Liu et al., 2014). Thus, in many respects, CT is a natural target for sustained study.

In Section 6, we discuss a number of technical issues related to such a study. A complementary aspect of our proposal is more pragmatic in nature. The Wabash National Study of Liberal Arts Education (Pascarella, 2007; Pascarella & Blaich, 2013) employed a heterogeneous sample of institutions. By contrast, we believe it can be advantageous to recruit – and to maintain the commitment of – a sample of institutions that share a common philosophy and similar approaches to education. In that regard, a consortium of liberal arts colleges (e.g., the Council of Independent Colleges) constitutes a potential 'universe' from which such a sample could be drawn.

However the sample is chosen, member institutions could collaborate in the design of the study so that they would derive direct benefits from its findings. As noted earlier, there is considerable evidence that many students do not demonstrate much, if any, growth in this domain. Thus, a comprehensive, longitudinal assessment that tracks growth would not only inform faculty about which particular populations of students need additional supports to thrive, but could also be used (for example) to evaluate interventions designed to promote CT. Finally, participants would be motivated, in part, by the hoped-for empirical validation of their instructional and co-curricular programs.

In that light, we propose that the initial study should be undertaken by a sample of schools from a particular consortium; namely, the Association of Jesuit Colleges and Universities (AJCU). Our choice is motivated, in part, by our location at a member institution and the expectation that the strong ties of that institution to other members of the AJCU will facilitate recruitment. On a different level, the study of CT should be of great interest to the AJCU: The Jesuit order has a centuries-long tradition of humanistic education that can be considered a forerunner of today's approaches to liberal education (O'Malley, 2015). Indeed, CT has always had a prominent role in Jesuit education. Moreover, the Jesuit tradition has embraced a more capacious conception of CT – one that encompasses not only analytic and synthetic thinking (in various forms), but also the

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

importance of confronting the challenges entailed in addressing ethical conundrums and moral dilemmas.

In view of their longstanding commitment to liberal education and to the essential role of CT, in many ways AJCU members constitute an ideal setting for conducting a systematic, sustained research program on the development of CT skills through the four years of college. The program should aim to document the extent and scope of CT development across a broad range of students, not only providing useful descriptions of students' trajectories, but also informing strategies on how to improve curricular and co-curricular offerings to better support CT development.

We also note that Jesuit colleges and universities span the globe and offer excellent opportunities to expand the research study beyond the U.S. At the same time, the AJCU members comprise a specialized sector of higher education and, quite appropriately, the question of the generalizability of the results arises. Generalizability is an important issue and is addressed in Section 6.8.

The article is organized as follows. The CT construct is introduced in the next section, followed by brief discussions of both liberal education and Jesuit education, with particular attention to their stances with respect to CT as an important learning outcome. The next two sections present considerations in the measurement of CT, with a focus on the utility of performance assessment. Then follows an extended presentation of the longitudinal study previewed above, with some details on a number of aspects of the research protocol that is proposed. The penultimate section discusses some of the methodological challenges, followed by a discussion section.

1. CRITICAL THINKING

Liu et al. (2014) provide a comprehensive review of the many definitions and frameworks for CT (See Table 1 in their paper for details). Summarizing their review, Liu et al. noted that there was general agreement that CT is a complex, multifaceted construct with five key, common facets: (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.

In the framework proposed by Braun et al. (2020), CT 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) (See also Shavelson et al., 2019). In carrying out a CT task, an individual typically engages in such activities as: (i) specifying or clarifying a problem; (ii) deciding what information is relevant to the problem; (iii) evaluating the trustworthiness of information; (iv) avoiding judgmental errors based on "fast thinking," biases, and stereotypes; (v) recognizing different perspectives and how they can reframe a situation; (vi) considering the consequences of alternative courses of actions; and (vii) communicating decisions and actions clearly and concisely. Braun et al. further suggest that CT includes (viii) dealing with dilemmas of ambiguity or conflict among (moral/ethical) principles and contradictory information. Oser and Biedermann (2020) argue that it is precisely this aspect of CT, which they label 'Critical Analysis,' that distinguishes CT from everyday logical reasoning.

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, 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

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

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

We maintain that this generic versus discipline-specific dichotomy obscures an underlying continuum. Performance assessments of CT (Braun et al., 2020) typically incorporate a number of sources of evidence. In a so-called generic CT assessment, there is an assumption that the respondent can provide an appropriate answer by just using the evidence provided, even though the challenge may well involve considerations from various disciplines. We term such an assessment "lightly grounded" in a disciplinary context, inasmuch as previous familiarity with some of the relevant issues may be helpful. Other assessments may be more explicit about drawing on a discipline, with some documents containing discipline-specific content (Minnameier & Hermkes, 2020). We label such assessments as "anchored" in a disciplinary context. Anchoring is a matter of degree, with some assessments relying very substantially on the content and methods of a discipline (e.g., a capstone project in an academic major). We describe such assessments as "deeply anchored" in the disciplinary context.

2. LIBERAL EDUCATION AND CRITICAL THINKING

Although CT is, or should be, a key learning outcome of any program of higher education, we focus here on programs that draw on traditions of liberal education. The AAC&U (2002) defined liberal education as:

> A philosophy of education that empowers individuals, liberates the mind from ignorance, and cultivates social responsibility. Characterized by challenging encounters with important issues, and more a way of studying than specific content, liberal education can occur at all types of colleges and universities. (p. 25)

Typically, liberal education involves a general education involving "broad exposure to multiple disciplines and forms the basis for developing important intellectual and civic capacities" as well as in depth study in one or more major fields. Expanding on this point, AAC&U (2020) asserts that:

> Through disciplinary study in general education and the majors, a solid undergraduate curriculum provides knowledge of human cultures and the physical and natural world. What matters for liberal education is that disciplinary study be focused by engagement with "big questions," both contemporary and enduring. Students also develop intellectual and practical skills—inquiry and analysis; critical and creative thinking; written and oral communication; teamwork and problem solving; quantitative, information, scientific, and technological literacies. (p. 9)

Finally, within the context of Jesuit higher education, Daley (1988) defined liberal education as follows:

> "Liberal" education means general education, education in values, education for wisdom rather than for marketability: not specialized training in the skills and information one needs for a career, but a process whereby one comes to know more fully the accomplishments and ideals of one's culture, in order to evaluate and redirect one's personal accomplishments and ideals. The goal of this general education, however one wants to define its content, is surely to work a kind of inner transformation: to stimulate a young mind to wonder at the world's beauty, to excitement at its complexity, to compassion at its vulnerability; to make that mind more deeply and reflectively aware of the ideals and values it is offered by its forebears, and to encourage a person, at the most adaptable and idealistic time of his life, to sift and organize those ideals for himself and commit himself fearlessly to what he sees to be good. (pp. 13-14)

Interestingly, the general framework for liberal education found in the U.S. can be traced back to classical Athens and, subsequently, to the rise of humanistic schools and colleges in 15th century

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

Europe (O'Malley, 2015). Nonetheless, during the last 150 years, liberal education in the European public tertiary sector was practically non-existent. It is only since 1990 that there has been a modest resurgence in establishing liberal arts colleges that are roughly modeled on U.S. institutions, though with some important differences (van der Wende, 2011).

In the view of van der Wende (2013), liberal education re-emerged in Europe and Asia in response to growing concerns about the quality of undergraduate education. More specifically, the massification of higher education around the world has led some European countries to adopt liberal arts education as a flexible, interdisciplinary alternative to the hyper-specialized, narrow curriculum that has characterized European higher education (van der Wende, 2011). Moreover, there is growing recognition in Europe of the need to improve students' generic skills—such as their writing and analytical skills—that are imperative for success in the modern economy (van der Wende, 2011). In China, the liberal arts are viewed as a means to spark innovation by fostering students' creativity (Boyle, 2019). Furthermore, the emphasis on moral development and social responsibility in liberal arts education is consistent with the Confucian belief systems in China (Cheng & Zhang, 2020).

Thus, the issues explicated here are generally relevant to institutions around the globe (Godwin & Altbach, 2016). Indeed, apart from the U.S., liberal arts education programs can be found in 57 countries, including Canada, India, Japan, Hong Kong, China, Australia, Netherlands, and Germany (Godwin, 2013; Godwin & Altbach, 2016).

It seems self-evident that CT in its various forms, is not only an essential learning outcome of a liberal education, but also is a capability that is intimately involved with achieving many of the other primary outcomes of liberal education. This is particularly the case of liberal education as it has been formulated and refined in Jesuit education – a subject to which we now turn.

3. JESUIT EDUCATION

Although liberal education is generally viewed through a secular lens, the goals of liberal education (e.g., understanding human culture and the natural world; intellectual and practical skills; personal and social responsibility), as articulated by AAC&U (2005), are in fact well-aligned with the goals of Jesuit education. O'Malley (2015) describes the purpose of Jesuit education as to:

> help the fly to fly out of the bottle, that is, to allow students to escape from the confines of their experience up to the present, to expand their awareness beyond the comfort zones of thinking in which they have grown up, to expose them to other cultures and to other modes of thought, to lift them beyond the quotidian. To help them escape from the bondage of unexamined assumptions and prejudices. To help them expand their consciousness and the areas in which they can dare to ask questions, not only in the areas in which their trade, discipline, or profession moves, but about life itself. (p. 28)

An educational journey that encourages students to move beyond their personal experiences and comfort zones; to examine prior unchallenged beliefs; and to expand their minds so as to consider new questions about the content of their courses and life more broadly surely fosters CT. Indeed, CT skills allow students to think through ambiguous, complicated real-world situations that do not necessarily have a clear-cut solution (i.e., "webs that are not neat geometrical patterns but are broken in places and often filled with knots and tangles" (O'Malley, 2015, p. 31)). Jesuit education not only fosters liberal education in general, but CT skills in particular.

Although liberal education today is not typically viewed from a Jesuit or other sectarian perspective, Morrill (2012) argues that religion should be a critical component of liberal education, as it provides students the tools needed for "interrogating our forms of life" (p. 6). However, academia has largely situated "questions of values and of religion a largely private matter that stand outside the kinds of evidence and argumentation that prevail in academic disciplines" (Morrill, 2012, p. 4).

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

Furthermore, Morrill argues that higher education's hesitation to make "broad value claims" has paradoxically made it more challenging to integrate some of the goals of liberal education into the curriculum, such as helping students think about their meaning and values and their larger civic responsibilities (p. 4). By comparison, the humanities have been utilized in Jesuit education as a mechanism for students to consider their larger meaning and purpose in the world in relation to their broader community, using experiential learning to promote reflection about social justice and inequality (Wortham et al., 2020).

4. MEASUREMENT OF CT

The Wabash National Study of Liberal Arts Education constituted an empirical investigation of the outcomes of liberal education over the four years of college (Pascarella & Blaich, 2013). In its first cycle, the study sampled students from 19 institutions of different types and followed them for four years. For this study, liberal education was operationally defined as comprising three key facets:

> 1. An institutional ethos and tradition that place a greater value on developing a set of intellectual arts than on developing professional or vocational skills. 2. Curricular and environmental structures that work in combination to create coherence and integrity in students' intellectual experiences. 3. An institutional ethos and tradition that place a strong value on student-student and student-faculty interactions both in and out of the classroom. (Blaich et al., 2004, p. 2)

Earlier we described CT as a complex, multi-faceted construct. However, most commercially available instruments employ only multiple-choice items that are generally recognized as inadequate to the task. As Liu et al. (2014) state:

> A major challenge in designing an assessment for critical thinking is to strike a balance between the assessment's authenticity and its psychometric quality. Most current assessments rely on multiple-choice items when measuring critical thinking. The advantages of such assessments lie in their objectivity (particularly with respect to scoring), efficiency, high reliability, and low cost. Typically, within the same amount of testing time, multiple-choice items are able to provide more information about what the test takers know as compared to constructed-response items (Lee et al., 2011). (p. 8)

In the above quote, 'authenticity' presumably refers both to 'face validity' and to 'construct representation.' However, the capability of multiple-choice items 'to provide more information' must be understood as referring to a limited number of facets of the construct – recognizing that some facets are refractory to measurement using selected response items. It is noteworthy, that the Wabash National Study (Pascarella & Blaich, 2013), one of the few rigorous studies of CT development, employed a multiple-choice instrument, the Critical Thinking Test of the Collegiate Assessment of Academic Proficiency (then marketed by the ACT). Thus, future research can build on the Wabash National Study by employing instruments that incorporate performance assessment.

Because (extended) performance assessments require greater respondent time and investments in scoring, there is considerable pressure to rely more (or solely) on multiple-choice (m-c) items. One argument is that typically one finds high correlations between scores on m-c items and scores on performance assessments (Klein et al., 2009). However, correlations are not sufficient to claim that students scoring high on an m-c assessment would meet a standard of achievement defined in terms of the full set of facets of the construct (See, for example, the VALUE scoring rubric for CT developed by the AAC&U (2009)). This point is especially germane in an instructional context, in which the goal is to have students achieve proficiency with respect to the standards set by the

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

faculty. Typically, those standards are framed (at least implicitly) with respect to multiple facets of the CT construct, including those not amenable to assessment by m-c items.

For example, Stassen et al. (2011) surveyed a large number of faculty and administrators at a large, flagship state university to obtain their characterizations of valued aspects of CT. The results were coded and clustered and then compared to the frameworks and/or instruments for a number of assessments. Two were standardized assessments offered by ACT and ETS, consisting solely of m-c items. One of their conclusions was: "Results demonstrate that the definitions reflected by standardized tests are more narrowly construed than those of the campus and leave dimensions of critical thinking unassessed" (p. 126).

This point brings to the fore a second tension, that between standardization and instructional relevance (Liu et al., 2014). In principle, administration of a standardized assessment of CT across multiple institutions should enable making informative comparisons among institutions. However, as Braun (2019) argues, the record demonstrates that there are numerous obstacles to actually conducting such comparisons and that investments could be more productively employed in strengthening the assessment of CT for instructional relevance. Further, he argues that in such an assessment program, well-designed performance assessments have an important role to play.

With these considerations in mind, we argue that the instrumentation for a comprehensive study of CT should complement objectively scored items with well-designed performance assessments. Collectively, the ensemble would be able to elicit evidence with respect to multiple facets of CT. However, these different item types should not be thrown together without care. A formal process, such as evidence-centered design (ECD) (Mislevy et al., 2003; Mislevy & G. Haertel, 2006) ought to be employed. ECD makes explicit the rationale for each step of the design process. This not only contributes to enhancing the quality of the instrumentation, but also builds in the documentation to support the validity argument. The ECD process results in an assessment 'blueprint' that provides explicit guidance on the types and numbers of items, as well as any auxiliary materials. The process is modified slightly in the case of an extended performance assessment, such as the one described in Braun et al. (2020). Since our proposal strongly emphasizes the importance of employing performance assessments, we now turn to a brief discussion of this form of assessment.

5. PERFORMANCE ASSESSMENT

Performance assessments "seek to emulate the context or conditions in which the intended knowledge and skills are actually applied" (AERA et al., 2014, p. 77). Thus, the challenge posed to the student by the assessment and the characteristics of the desired response are meant not only to be similar to what might be observed in real-world settings, but also to provide evidence to support intended interpretations and actions. Although performance assessment falls under the general rubric of constructed response, an elaborate simulation task can be readily distinguished from such exercises as fill-in-the-blank or carrying out a decontextualized numerical computation.

Adjectives such as "authenticity," "fidelity," and "transparency" are often used in conjunction with performance assessments. However, these adjectives must be interpreted both in light of the purpose of the assessment and the setting in which it is administered. With some exceptions (e.g., sports or dance competitions), a performance assessment can only properly reflect certain aspects of a real-world setting, while short-changing or neglecting others. Hence, validating the use of the assessment requires judgment based on both theoretical analysis and empirical evidence. As Messick (1994) argues, a performance assessment is only a promissory note for the elicitation of credible evidence regarding student proficiency. In particular, performance assessments used as part of an instructional program and performance assessments embedded in a (high-stakes) summative instrument demand very different validation strategies.

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

In the context of higher education, well-designed performance assessments of CT can elicit evidence of students' proficiencies with respect to a number of facets. In an instructional context, the evaluation of their responses, guided by an elaborated scoring rubric, yields feedback that can provide useful information as to where further effort is required, as well as what kinds of tasks could be employed as part of instruction.

With regard to CT, there are a number of extant CT performance assessments. Perhaps the best known is the Collegiate Learning Assessment (CLA) that, in its original form, addressed different facets of CT (Klein et al., 2007; Zahner, 2013). The Reflective Judgment Interview (Kitchener & King, 1985) is a semi-structured interview based on a model of reflective judgment (King & Kitchener, 2002). The interview protocol has been adapted into an assessment that is administered on a computer. The Critical Reasoning Assessment (*CRA*; Anghel et al., in press) presents respondents with a moral-cognitive question (e.g., do people succeed due to effort or privilege?). Respondents then answer a set of seven open-ended questions, requiring them to present their opinions and the evidence they used to support it, as well as evidence that might support alternative views. Responses are evaluated using a highly refined, analytic scoring rubric (i.e., a rubric with multiple scoring categories).

Finally, under the auspices of the International Performance Assessment of Learning (iPAL) consortium, an explicit assessment framework is used to guide development of complex performance assessments (Braun et al., 2020). The assessment includes four main components: (1) The storyline describes a carefully curated version of a complex, real-world situation. (2) The challenge frames the task to be accomplished, with reference to the documents provided and with varying degrees of scaffolding. (3) A portfolio of documents in a range of formats (e.g., reports, charts, blogs, twitter threads) is drawn from multiple sources chosen to reflect different levels of relevance, trustworthiness, and susceptibility to bias. (4) The scoring rubric comprises a set of rating scales each linked to a facet of the CT construct.

6. THE PROPOSAL

We propose that a sample of U.S. Jesuit colleges and universities undertake an extended, comprehensive study of the development of CT among their students. Such a study would provide invaluable evidence regarding trajectories of CT development and levels of proficiency. The promise of building an evidence base for their success (or lack of success) in supporting CT development, as well as for evaluating the efficacy of both curricular and co-curricular programs, should lead to enrolling a number of colleges in the proposed study. Lastly, given their decision to enroll in a Jesuit institution, one can surmise that the students recruited for the study will be more committed than the general college population to the study's broader aims, resulting in lower attrition rates.

Perhaps the most obvious question at this juncture is: Why Jesuit schools? We believe there are a number of factors that make Jesuit schools a natural setting for such a study. First, as noted above, the aims of Jesuit education consider CT both as a learning outcome in its own right and as essential to accomplishing other outcomes – during the college years and beyond. Those outcomes encompass both personal development and preparation for the world of work. As O'Malley (2015, p. 23) notes: "The Jesuits were among the educators who did not see an unbridgeable gap between professional and humanistic training." Consequently, various forms of analytic thinking have always loomed large in the Jesuit curriculum.

Speaking of the goals of higher education, Morrill (2012) argued "Just as we aim in universities to teach people how to think, so we can legitimately aspire to teach students how to value and encourage and enable them to develop an internalized critical apparatus for making choices among

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

values and forms of life" (p. 6). Certainly, Jesuit education does not shy away from the important questions and is comfortable with raising issues that have moral/ethical aspects. Thus, confronting students with scenarios that involve, among other things, ethical dilemmas and moral challenges, would be consistent with the Jesuit educational ethos. Clearly, there are parallels between helping students discern their life purpose and CT more broadly.

Finally, from a pragmatic point of view, Jesuit colleges are private, independent institutions with the flexibility to participate in research studies that align with their institutional missions. Moreover, although the vast majority of liberal arts programs are located in the United States (Godwin, 2013), it is reasonable to expect that subsequent studies could include Jesuit and other liberal arts institutions located outside the U.S.

One of the few rigorous, longitudinal studies of college student development was conducted under the auspices of the Wabash National Study of Liberal Arts Education (Pascarella & Blaich, 2013). The technical reports generated through the course of the study convey very graphically the complexity of such an undertaking, and provide a useful blueprint for the design of the proposed study. In the sections that follow, we explicate some of the main considerations in the planning: funding, recruitment, attrition, data collection design, instrumentation, analysis and validation, and pragmatics.

6.1 FUNDING

There are considerable costs associated with conducting a longitudinal study that spans (say) four years, encompassing a planning year, two years of data collection, and one year for analysis and reporting. A combination of foundation funding and in-kind contributions by participating institutions is the most likely mechanism to secure sufficient support for the project. With such studies, patience among funders and stakeholders, must be assiduously cultivated and maintained.

6.2 RECRUITMENT

Recruitment operates at two levels: institutional and student (within institution). Ideally, participating institutions would be generally representative of the membership of the Association of Jesuit Colleges and Universities (AJCU) with regard to size and geographic region. In practice, however, this would be a convenience sample that includes institutions with stronger commitments to CT development.

Student recruitment is more challenging, as students cannot be compelled to participate. Recruitment will likely require an appeal to contribute to the greater good, as well as incentives of some type. Maintaining a rough comparability across institutional samples by drawing simple random samples (or with sampling proportions specific to institutional demographics) is desirable but not absolutely necessary. On the other hand, one could choose to oversample specific groups defined by combinations of gender, race/ethnicity, major, or other student characteristics. This would permit more accurate sub-group estimates, as well as more informative institutional comparisons. For smaller institutions, a full census may be more practicable.

6.3. ATTRITION

Beyond initial recruitment, experience shows that maintaining participation in a longitudinal study is even more difficult. Although institutional-level attrition is not likely, it is possible – and very problematic. One way to mitigate this possibility is to ensure that the initial commitment is at the institutional level (i.e., via the institution's senior leadership team) and not the purview of a single champion. With regard to students, attrition is to be expected. In the Wabash study, attrition over

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

the four years was approximately 47 percent (Pascarella et al., 2013). In a two-year study, it should be possible to reduce attrition to below 25 percent.

Beyond the counts, the reasons for students discontinuing participation are germane to data analysis and inference. Especially concerning is if student attrition is related to CT, the target of measurement. For example, students with lower levels of CT may be less inclined to continue with the study or may even leave school. Although there are some strategies to mitigate the ensuing bias, they cannot fully compensate for the sample loss.

6.4 DATA COLLECTION DESIGN

As noted above, longitudinal studies are both difficult and costly to carry out. In estimating growth, the Council for Aid to Education, sponsor of the CLA, has employed cross-sectional designs with simultaneous administration of the CLA to first-year and fourth-year students. The results are then statistically adjusted to account for prior (academic) differences between cohorts, as well as for attrition (Klein et al., 2007). However, this approach has garnered a number of criticisms (Banta & Pike, 2007). Alternative approaches requiring parallel two-year longitudinal designs have been proposed but not implemented. These and other designs should be examined in light of a careful analysis of advantages and disadvantages of each.

6.5 INSTRUMENTATION

Demographics and other academic-related information could be obtained from school databases, with appropriate attention to privacy concerns. In order to capture the full range of the CT construct, while achieving satisfactory levels of reliability, it will be necessary to develop a web-based instrument that comprises items with different formats – from multiple choice items to extended performance assessments. With respect to the latter, both the CRA and the iPAL performance tasks could be utilized, at no charge, in such a study (Anghel et al., in press; Braun et al., 2020; Shavelson et al., 2019). Ongoing collaborations among iPAL participants offer a proof of concept regarding the feasibility of developing a cross-national assessment. For instance, iPAL performance tasks have successfully been translated and adapted from English to Spanish and German to English. Exemplar scoring rubrics are available in German (Zlatkin-Troitschanskaia et al., 2019) and Spanish (Mejía et al., 2019).

Whichever performance tasks are employed, they would likely be complemented by a set of objectively scored items that draw on comparable scenarios and target specific facets of the CT construct. The CT tasks will be "lightly grounded" in various academic disciplines but will not require specific subject matter expertise. Additional instruments could include those measuring related constructs such as moral development, moral agency, and purpose in life, which might provide added incentive for Jesuit institutions to participate given the centrality of these constructs to their missions.

It is essential to complement the set of quantitative assessments with more qualitative measures employing a mixture of surveys and focus groups. The results would provide a rich context for interpreting the outcomes of the quantitative assessments, as well as being of interest in their own right.

Additional information collected about students' engagement in co-curricular programs (e.g., service-learning) and extra-curricular activities (e.g., athletics, theatre) will supplement the data about students' academic development. Jesuit institutions have long emphasized these outside-of-class experiences as integral to educating the whole-person (O'Malley, 2015).

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

6.6. ANALYSIS AND VALIDATION

Analysis will depend on the specific research questions posed and the data collected. Standard approaches should provide a baseline that subsequent, more sophisticated analyses, can build upon. Of particular interest are the CT score trajectories at the student and sub-population levels, complemented by their relationships to student- and sub-population-level characteristics. Latent class analyses may also prove useful for providing holistic profiles of students' skills. Specific instructional interventions could then be designed based on these profiles.

Validation strategies must be built into the study design. Following the model proposed by Kane (2013), an interpretation/use argument should be proposed at the outset and the accompanying validation argument devised. The latter will specify the characteristics of the data that must be collected to support the validity argument. In addition, patterns of relationships among the quantitative measures, as well as between those measures and the data collected through qualitative methods, can all be used to support the validation argument.

6.7 PRAGMATICS

In addition to ensuring that the research is supported by the institution's senior leadership team, administering a study of even modest scope will likely involve hiring a part-time project managers and several graduate research assistants, providing a valuable professional development opportunity. Furthermore, the research will likely involve collaboration with each institution's institutional research/assessment director. A small stipend could be provided to this staff member in recognition of the extra work added to their portfolio.

6.8 GENERALIZABILITY

As noted above, there are clear advantages to conducting such a study within a relatively homogeneous sample of institutions. However, as a reviewer pointed out, this design markedly reduces the generalizability of the findings. We agree – nonetheless, we believe that the long-term goals of this effort are best served by carrying out a series of studies within each of a set of homogeneous institutions. The findings of a study based on a representative sample of Jesuit colleges could be plausibly (though cautiously) extrapolated to the full set of 27 member institutions. For example, we could extract institution-level data from the National Center for Education Statistics (U.S.) about all AJCU institutions, and then evaluate how closely our sample parallels the overall set on key institutional characteristics (e.g., student to faculty ratio, institution resources).

Although AJCU institutions comprise a distinct sector within higher education, they have much in common with non-sectarian colleges that have a strong liberal education ethos. Thus, the proposed study could serve as a baseline for future studies conducted within other sectors of higher education. Similarities and differences in the findings would be informative in their own right.

We note that generalizability across institutions is even challenging in studies with varying institutional types. For example, the Wabash National Study enrolled three cohorts of students/institutions in 2006, 2007, and 2008, respectively (Center of Inquiry at Wabash College, n.d.). The first cohort included 19 institutions of four different types: liberal arts colleges, research universities, regional institutions, and community colleges (Pascarella & Blaich, 2013). The numbers of each type were too small to make general statements. Moreover, as each sub-sample was a convenience sample from the respective sector, the usual, comparative statistical inference procedures were not available. Further, the second and third cohorts included seven and 26 institutions, respectively (Center of Inquiry at Wabash College, n.d.). Both cohorts included liberal

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

arts colleges, research universities, and regional institutions, and the third cohort also included a community college. Since some institutions were represented in multiple cohorts, 49 unique institutions were represented across all three cohorts. Nonetheless, effectively only qualitative comparisons across institutional sectors were possible.

7. METHODOLOGICAL CHALLENGES

Among students who participate in the assessment, absent some other meaningful objective (to the student), it is likely that many students will participate with less than maximal effort or even drop out of the study. This can be particularly problematic with tasks involving performance assessments that require elaborated responses and for which the stakes are relatively low for students (Lane & Stone, 2006; Shavelson, 2013). One possible solution is providing multiple tasks and allowing students to select the one that most interests them (Lane & Stone, 2006). Payments, say in the form of gift cards (either through a lottery or universally applied) can be effective, especially if linked to the level of performance (Braun et al., 2011).

It bears mentioning that there are also a number of measurement issues to be addressed. Because of practical time constraints, as well as the need to minimize respondent burden, the number of performance assessments administered at each stage is likely to be one or, at most, two. Consequently, individual-level score reliability is typically lower than one would prefer (Davey et al., 2015; Lane & Stone, 2006; Shavelson et al., 1993). However, the primary focus of the study will be group-level scores, with higher levels of reliability.

The reliability of scores on performance tasks is also impacted by disagreements among scorers. Obtaining a high level of inter-rater reliability can be particularly challenging when evaluating responses to performance tasks (Braun, 2019; Lane & Stone, 2006; Shavelson, 2013). Several rater behaviors can interfere with reliable scoring, due to variability in behavior within and across raters (Zhang, 2013). For examples, raters might differ from one another in how they interpret the scoring categories and individual raters might change their scoring criteria over time (i.e., "rater drift") (Bejar, 2012; Zhang, 2013). Moreover, in their interviews with raters, Zlatkin-Troitschanskaia et al. (2019) found that raters had difficulty distinguishing among multiple scoring dimensions. Additional construct-irrelevant factors that can impact raters' assessment of student performance include the length of the response and mechanical errors in the response (Lane & Stone, 2006). Therefore, it is essential to carefully train raters; to closely monitor raters' performance; and to provide raters feedback about their performance and additional training as needed (Braun, 2019; McClellan, 2010; Wolfe & Song, 2016). Finally, credible comparisons across occasions depend on a reasonable level of comparable difficulty between assessments and this is challenging with performance assessments.

For some purposes, it may be important to demonstrate the efficacy of a particular instructional program. In that case, one may want to go beyond demonstrating that progress has occurred; that is, the claim of efficacy must be examined by comparing the progress of students in the program to that of students in some other (control) condition. Choosing an appropriate control, designing the study and analyzing the results can be a substantial undertaking requiring an additional commitment of time and funds.

8. FINAL THOUGHTS

In agreement with Katz (2008), we believe that the defense of liberal education has fallen far short of what could – and should – be done. Hatch (2012) cautions, "We must not underestimate the danger that humanistic inquiry will wither into irrelevancy" (p. 6). The time is ripe for higher education institutions to respond to criticisms regarding student learning outcomes. The COVID-

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

19 pandemic will likely have long-term implications for college students' labor market experiences. Organizations will need employees who are capable of quickly adapting to changing workforce needs. CT skills will prove invaluable in equipping college students to solve the many problems facing the world today (Finley, 2021). Moreover, all adults will face multiple responsibilities in dealing with the increasing complexity of modern life, many of which will call on various aspects of CT (Hacker, 2019).

We have argued that a rigorous longitudinal assessment of students' CT skill development is much needed. Subsequently, specific instructional or co-curricular interventions could be developed and evaluated to support students' CT growth. Jesuit institutions should be a welcoming setting to conduct an initial longitudinal study, since their humanistic values are well-aligned with CT skill development. We recognize that our proposed study will not be without its challenges, including those related to student recruitment and retention, as well as students' motivation to perform well on low-stakes assessments. Some of these challenges can be mitigated through meaningful incentives and a high degree of institutional support for the research. Nonetheless, recognition of these challenges should not be an excuse to refrain from action – rather, they are the basis for developing a realistic assessment of needed resources, robust designs, and appropriate analytic procedures.

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anghel, E., Braun, H. I., Friedman, A. A., & Baez-Cruz, M. (in press). College students' critical thinking: Assessment and interpretation. *Journal of Higher Education Theory and Practice*.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Association of American Colleges and Universities. (2002). *Greater expectations: A new vision for learning as a nation goes to college*. Washington, DC. Retrieved from https://www.aacu.org/sites/default/files/files/publications/GreaterExpectations.pdf

Association of American Colleges and Universities. (2005). *Liberal education outcomes: A preliminary report on student achievement in college*. Retrieved from https://www.aacu.org/sites/default/files/files/LEAP/LEAP_Report_2005.pdf

Association of American Colleges and Universities. (2009). *Critical thinking VALUE rubric*. Retrieved from https://www.aacu.org/value/rubrics/critical-thinking

Association of American Colleges and Universities. (2020). *What liberal education looks like: What it is, who it's for, & where it happens*. Retrieved from https://www.aacu.org/publications-research/publications/what-liberal-education-looks-what-it-who-it%E2%80%99s-and-where-it

Banta, T., & Pike, G.R. (2007). Revisiting the blind alley of value-added. *Assessment Update, 19*(1), 1-2, 14-15.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31* (3), 2-9.

Blaich, C., Bost, A., Chan, E., & Lynch, R. (2004). *Defining liberal arts education* (Unpublished manuscript). Center of Inquiry in the Liberal Arts at Wabash College.

Bok, D. (2006). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton University Press.

Boston College. (2007). *The journey into adulthood: Understanding student formation*. Retrieved from https://www.bc.edu/content/dam/files/offices/mission/pdf1/umm1.pdf

Boyle, M. E. (2019). Global liberal education: Theorizing emergence and variability. *Research in Comparative and International Education, 14*(2), 231-248.

Braun, H. (2019). Performance assessment and standardization in higher education: A problematic conjunction? *British Journal of Educational Psychology, 89*(3), 429–440. doi: 10.1111/bjep.12274

Braun, H. I., Kirsch, I., & Yamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309–2344.

Braun, H. I., Shavelson, R.J., Zlatkin-Troitschanskaia, O., & Borowiec, K. (2020). Performance assessment of critical thinking: Conceptualization, design, and implementation. *Frontiers in Education, 5*, 156. doi: 10.3389/feduc.2020.00156

Center of Inquiry at Wabash College. (n.d.). *Wabash National Study participating institutions*. Retrieved from https://centerofinquiry.org/wabash-national-study-participants/

Cheng, B., & Zhang, D. (2020). Cultivating citizens with Confucian cosmopolitanism: Defining the purpose of liberal arts education in the Asian context. *Frontiers of Education in China, 15*(4), 564-587.

Daley, B. (1988). "Splendor and wonder": Ignatian mysticism and the ideals of liberal education. In W. J. O'Brien (Ed.), *Splendor and wonder: Jesuit character, Georgetown spirit, and liberal education* (pp. 1-22). Georgetown University Press.

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

Delbanco, A. (2012). *College: What it was, is, and should be*. Princeton University Press.

Finley, A. (2021). *How college contributes to workforce success: Employer views on what matters most*. Washington, DC: Association of American Colleges and Universities. Retrieved from https://www.aacu.org/sites/default/files/files/research/AACUEmployerReport2021.pdf

Godwin, K. A. (2013). *The global emergence of liberal education: A comparative and exploratory study* [Doctoral dissertation, Boston College].

Godwin, K. A., & Altbach, P. G. (2016). A historical and global perspective on liberal arts education: What was, what is, and what will be. *International Journal of Chinese Education, 5*(1), 5-22.

Grinnell College. (n.d.). *Our mission*. Retrieved from https://www.grinnell.edu/about/at-a-glance/mission

Hacker, J. S. (2019). *The great risk shift: The new economic insecurity and the decline of the American dream*. Oxford University Press.

Hatch, N. (2012, Nov. 8). Hope and challenge in the middle ground [Speech transcript]. In E. Owens (Chair), *Boston College Symposium on Religion and the Liberal Aims of Higher Education* [Symposium]. Boisi Center for Religion and American Public Life, Boston College, Chestnut Hill, MA. Retrieved from
https://www.bc.edu/content/dam/files/centers/boisi/pdf/f12/RLE%20Hatch%20Keynote.pdf

Indiana University (2019). *FSSE 2019 frequencies: FSSE 2019 aggregate*. Retrieved from http://fsse.indiana.edu/pdf/FSSE_IR_2019/summary_tables/FSSE19_Frequencies_(FSSE_2019).pdf

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

Katz, S. N. (2008, May 23). Taking the true measure of a liberal education. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/weekly/v54/i37/37a03201.htm

King, P. M., & Kitchener, K. S. (2002). The reflective judgment model: Twenty years of epistemic cognition. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 37–61). Lawrence Erlbaum Associates, Inc.

Kitchener, K. S., & King, P. M. (1985). *The reflective judgment scoring manual* (Unpublished manuscript).

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review, 31*(5), 415–439.

Klein, S. C., Liu, O. L. E., Sconing, J. A., Bolus, R. C., Bridgeman, B. E., Kugelmass, H. C., ... & Steedle, J. C. (2009). *Test validity study (TVS) report*. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.173.3647&rep=rep1&type=pdf

Kronman, A. (2008). *Education's end: Why our colleges and universities have given up on the meaning of life*. Yale University Press.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 387-432). Rowman & Littlefield Publishers.

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 RR–14-10). Educational Testing Service. doi:10.1002/ets2.12009

McClellan, C. A. (2010, Feb.). Constructed response scoring doing it right. *R&D Connections*, 13.

Mejía, A., Mariño, J. P., & Molina, A. (2019). Incorporating perspective analysis into critical thinking performance assessments. *British Journal of Educational Psychology, 89*(3), 456-467.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23. doi: 10.3102/0013189x023002013

Minnameier, G., & Hermkes, R. (2020). Learning to fly through informational turbulence: Critical thinking and the case of the minimum wage. *Frontiers in Education, 5*, 573020. doi: 10.3389/feduc.2020.573020

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i-29.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20. doi: 10.1111/j.1745-3992.2006.00075.x

Morrill, R. (2012, Nov. 9). Religion and a larger vision for liberal education [Speech transcript]. In E. Owens (Chair), *Boston College Symposium on Religion and the Liberal Aims of Higher Education* [Symposium]. Boisi Center for Religion and American Public Life, Boston College, Chestnut Hill, MA. Retrieved from https://www.bc.edu/content/dam/files/centers/boisi/pdf/f12/RLE%20Morrill%20Keynote.pdf

O'Malley, J. W. (2015). Jesuit schools and the humanities yesterday and today. *Studies in the Spirituality of Jesuits, 47*(1), 1-34.

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

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

Pascarella, E. T. (2007). *Methodological report for Wabash National Study of Liberal Arts Education*. Iowa City, IA: Center for Research on Undergraduate Education. Retrieved from https://centerofinquiry.org/wp-content/uploads/2017/04/WNSLAE_Research_Methods_March_2008.pdf

Pascarella, E. T., & Blaich, C. (2013). Lessons from the Wabash National Study of Liberal Arts Educaton. *Change: The Magazine of Higher Learning, 45*(2), 6–15. https://doi.org/10.1080/00091383.2013.764257

Pascarella, E. T., Wang, J. S., Trolian, T. L., & Blaich, C. (2013). How the instructional and learning environments of liberal arts colleges enhance cognitive development. *Higher Education, 66*(5), 569–583. https://doi.org/10.1007/s10734-013-9622-z

Pratt, M. (2016). *AUC 21: Our community's vision for excellence, diversity and global citizenship*. Retrieved from https://www.auc.nl/about-auc/mission-and-values/auc21/auc21.html

Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist, 48*(2), 73-86.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. P. (2019). Assessment of university students' critical thinking: Next generation performance assessment. *International Journal of Testing, 19*(4), 337-362.

Spellings, M. (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington, D.C.: U.S. Department of Education. Retrieved from

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

https://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/pre-pub-report.pdf

Stassen, M. L., Herrington, A., & Henderson, L. (2011). Defining critical thinking in higher education: Determining assessment fit. In J. E. Miller & J. E. Groccia (Eds.), *To improve the academy: Resources for faculty, instructional, and organizational development* (Vol. 30) (pp. 126-141). John Wiley & Sons.

Van der Wende, M. (2011). The emergence of liberal arts and sciences education in Europe: A comparative perspective. *Higher Education Policy, 24*(2), 233-253.

Van der Wende, M. (2013). Trends towards global excellence in undergraduate education: Taking the liberal arts experience into the 21st century. *International Journal of Chinese Education, 2*(2), 289-307.

Wheeler, P., & Haertel, G. D. (1993). *Resource handbook on performance assessment and measurement: A tool for students, practitioners, and policymakers*. Owl Press.

Wolfe, E.W. & Song, T. (2016). Methods for monitoring and document rating quality. In H. Jiao & R.W. Lissitz (Eds.), *The next generation of testing* (pp. 107-142). Information Age Publishing.

Wortham, S., Love-Jones, R., Peters, W., Morris, S., & García-Huidobro, J. C. (2020). Educating for comprehensive well-being. *ECNU Review of Education, 3*(3), 406-436.

Zahner, D. (2013). *Reliability and validity–CLA+*. New York: Council for Aid to Education.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections, 21*, 1-11.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., & Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *British Journal of Educational Psychology, 89*(3), 468-484.

DOI: https://doi.org/10.15366/jospoe2021.13.006
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 125-144

ABOUT THE AUTHORS

Henry Braun

Henry Braun is the Boisi Professor Education and Public Policy in the Lynch School of Education & Human Development at Boston College. He holds a Ph.D. in mathematical statistics from Stanford University. After serving as an assistant professor of statistics at Princeton University, he joined the Educational Testing Service in 1979, where he served as vice-president for research management from 1990 to 1999. He held the title of distinguished presidential appointee from 1999 until his retirement in 2006 when he moved to Boston College. A fellow of the American Statistical Association and of the AERA, he is an elected member of the National Academy of Education. Dr. Braun is a co-recipient of the 1986 Palmer O. Johnson Award of the AERA and a co-recipient of the NCME's 1999 Award for Outstanding Technical Contribution to the Field of Educational Measurement. He received a T.J. Alexander Fellowship from the OECD (2014) and the 2018 Robert L. Linn Distinguished Address award from AERA (Div. D). His interests include the analysis of large-scale assessment survey data and social policies related to educational opportunity, school and teacher accountability, higher education learning outcomes, and the role of testing in education policy.

Contact information: Boston College Lynch School of Education and Human Development, 140 Commonwealth Avenue, Chestnut Hill, MA, henry.braun@bc.edu

Katrina Borowiec

Katrina Borowiec is a doctoral candidate in Measurement, Evaluation, Statistics, and Assessment at Boston College. Prior to her doctoral studies, she worked in institutional research at Mount Holyoke College in South Hadley, Massachusetts (USA). Her research interests include college students' experiences, academic outcomes, and well-being; psychosocial scale development; and measurement invariance.

Contact information: Boston College Lynch School of Education and Human Development, 140 Commonwealth Avenue, Chestnut Hill, MA, Katrina.Borowiec@bc.edu