Whiteacre, K (2024). Reconsidering the Direct Vs. Indirect Evidence Dichotomy. *Intersection: A journal at the intersection of assessment and learning, Association for the Assessment of Learning in Higher Education Conference Proceedings, 5*(3), 13-20.

Reconsidering the Direct Vs. Indirect Evidence Dichotomy

Kevin Whiteacre, Ph.D.

Author Note

Kevin Whiteacre ORCID iD: https://orcid.org/0009-0000-1273-7104

The author has no conflicts of interest to disclose

Intersection: A Journal at the Intersection of Assessment and Learning
AALHE Conference Proceedings 2024, Volume 5 Issue 3

Abstract: High quality student outcome measures provide the foundation for effective learning assessment. These outcomes can be organized into different categories, such as skills or habits of mind, and evidence types, such as direct or indirect. Prevailing models of assessment, however, focus only on the distinction between direct and indirect evidence without incorporating the outcome categories. Such a binary model of evidence can be limiting, privileges measures of skills and knowledge over habits of mind and values, and risks excluding validated psychometric tools as a source of measurement. An argument is made for a new 2x2 outcome matrix which incorporates outcome category and evidence type as a possible alternative to the dominant direct vs. indirect dichotomy.

Keywords: *psychometrics, direct evidence, indirect evidence, learning outcomes*

Organizing Outcomes and Measures

Student learning outcomes (SLOs) can be usefully organized into different categories. Ewell (1987), for example, distinguishes four "basic dimensions" of outcome assessment: 1) Knowledge, emphasizing cognitive content, 2) Skills, emphasizing application, 3) Attitudes and Values, such as tolerance for diversity and empathy, and 4) Behavioral, "a manifestation of the other three outcomes," which includes employment, continued education, and salaries. Suskie's (2009) "framework" follows a similar model, minus the Behavioral dimension: 1) Knowledge and conceptual understanding, 2) Thinking and other skills, and 3) Attitudes, values and dispositions.

The operational measures or methods for tapping into those various outcomes can also be helpfully divided into categories, the most prevalent of which is the distinction between direct and indirect measures. "*Direct* evidence of student learning is the kind that would convince a skeptic that students indeed have learned what they need… *Indirect* evidence is less compelling evidence of student learning" (Suskie, 2018; italics in original). "...[W]hen selecting specific assessment techniques, activities that directly assess learning must be included if the assessment is to have credibility" (Palomba & Banta, 1999, p. 11-12).

Examples of direct measures include local and standardized exams, portfolios, presentations, papers, and other assignments generally assessed via rubrics (see Table 1 for common examples) (Palomba & Banta, 1999; Suskie 2009; Suskie, 2018). Indirect measures involve course grades, alumni surveys,

completion of internships and service learning, and student surveys and course evaluations (Palomba & Banta, 1999; Suskie 2009; Suskie, 2018).

Table 1

The Binary Model of Direct vs. Indirect Evidence

Direct	**Indirect**
-Standard tests -Some rubrics -Portfolios -Student reflections on values and attitudes	-Course evaluations -Student and alumni surveys -Internship completion -Service learning participation -Employment placement

As the definitions above suggest, direct measures are generally considered superior to indirect measures. Indeed, at the author's institution, annual learning outcome reports must identify whether each measure is direct or indirect. The provided one-page guide with examples states, "Indirect evidence provides signs that students are probably learning, but the evidence of exactly what they are learning is *less clear and less convincing.*" Yes, *both* the *italics* and underline are in the original to leave no doubt.

Unfortunately, the current binary distinction between direct and indirect evidence types often fails to account for the SLO category type, in particular attitudes, values and dispositions. Since attitudes and values are not readily assessed through the current definition of direct evidence, which is accepted as superior to indirect evidence, they are implicitly, if not purposefully, relegated to the margins of assessment. The issue taken up here is not whether direct evidence is superior to indirect, but whether the binary distinction might be limiting because: 1) It often assumes the validity of direct measures; 2) It implicitly diminishes attitudinal outcomes that are not as readily quantified through the accepted measures of direct evidence; and 3) In practice, if not intent, it excludes an important measurement: validated psychometric instruments, or at least writes them off as simply indirect evidence due to their reliance on surveys for data collection.

Over-Esteeming Tests and Rubrics as "Direct Evidence"

Citing Miselvy and colleagues' (2012) characterization of the *standard assessment paradigm* (SAP), Swiecki, et al. (2022) note several problems with relying on direct evidence like multiple-choice questions, essays, and short answer questions, to infer student knowledge and learning. Such measures may only provide discrete snapshots of performance with no comparison over time. They may be uniform, failing to account for the particular knowledge skills and backgrounds of participants, and they may be inauthentic, "adhering to the culture of schooling rather than the cultures schooling is

designed to prepare students to enter" (Swiecki, et al., 2022, p 1). Human scoring also raises a concern with reliability (Rhodes & Bergeron, 2017).

Ewell and Jones (1991) emphasize that "all indicators of education achievement are in some sense indirect" (p 6; emphasis in original). Learning outcomes are constructs, and even skills and knowledge tests are ultimately samples of the constructs used to infer learning. Content underrepresentation, therefore, poses another threat to the validity of most direct measures if not adequately tested and validated (Rhodes & Bergeron, 2017). Yet, direct evidence is often treated as if, as implied in the title "direct," they directly and concretely measure the whole construct validly and reliably. At the very least, these direct measures are bestowed a benefit of the doubt not granted to psychometric measures, conceived by many as mere indirect measures due to the reliance on surveys. This approach risks an uncritical acceptance of direct measures by dint of being "direct," even though they too, in the end, are inferences based on incomplete information.

Diminishing Values and Attitudes as an Outcome

> "There have been many changes and challenges [in American higher education] during my time as a college president. But one important thing hasn't changed: the value of a college education and its ability to transform students' lives." Marvin Krislov (2017)

The much-needed focus on and guidance for cleaning up vague and poorly measured SLOs with specific concrete measurable outcomes has left behind some very valuable SLOs, especially those associated with attitudes and values, such as social responsibility, openness to experience, or what might be called personal growth or character. This does not seem to be an intentional disregard as much as the unintended byproduct of an unyielding focus on a limited view of direct evidence, generally associated with skills and knowledge, due in part to the measures predominantly used.

Many university missions, and thus presumably their learning outcomes, however, explicitly invoke values, attitudes, or dispositions. Indeed, they underlie the liberal arts project, such as Brown University's mission to prepare students to "discharge the offices of life with usefulness and reputation" or William and Mary's to cultivate "global citizens equipped for lives of meaning and distinction." Unfortunately, these lofty outcomes aimed at "transforming students' lives," do not easily lend themselves to the knowledge and skills-based tests and rubrics favored by current distinctions between direct and indirect evidence. As a consequence, such outcomes, arguably the core value proposition of a liberal arts education, become secondary to basic knowledge or skill outcomes, more easily measured through the predominant direct outcomes of test and rubric scores[1].

Within universities, values and attitudes are especially important to the missions of their colleges of arts and sciences, but they remain difficult to nail down through current conceptions of direct

[1] One might even extend this logic to surmise the rise of "skills-based" credentialing has been fostered by the same issues presented by the direct/indirect view of measurement.

evidence. Purdue's College of Liberal Arts seeks cultivate "a capacity for openness, subtlety, and nuance as well as to nurture a passion for ideas and an appreciation of historical context…." Try assessing that with a rubric. The University of South Carolina's CAS value statement within its Mission likewise states "A liberal arts education prepares individuals to face an increasingly complex, diverse, and changing world with open, nimble minds and expansive, humane sympathies." Finally, the University of San Francisco's College of Arts and Sciences mission states "The College of Arts and Sciences offers wide variety of educational options imbued with a mission to provide not only the knowledge and skills needed to succeed as persons and professionals, but also the values and sensitivity necessary to be men and women for others." While some of these outcomes can potentially be tapped through portfolios and maybe student reflections, they pose a difficult (and cumbersome) challenge to do so validly, reliably, and efficiently.

Having nimble minds, a passion for ideas, capacity for openness and subtlety, and expansive, human sympathies–doubtless strong indicators of a transformed life–simply do not mesh with the current direct vs. indirect dichotomy and their accepted measurements. Either such aspirational transformative goals must be rejected as vague or too difficult to directly measure, or our notion of what constitutes valid direct evidence needs revision. Perhaps that revision can usefully integrate SLO type to inform the appropriate measurement.

A New Model

Validated Psychometric Tools

Many of the "transformative" outcomes of a liberal arts education can be tapped effectively through psychometric instruments explicitly designed to measure them. A capacity for nuance? What about the Ambiguity Tolerance scale, MAT-50, or Uncertainty Tolerance (Furnham & Ribchester, 1995; Paralkar & Knutson, 2023)? A religious institution might seek to assess student progress on Spirituality, Intrinsic Religious Motivation, or Religious Coping (Pargament, Feuille, & Burdzy, 2011) among its matriculating students. A university nurturing world citizens might usefully measure student scores on a validated Global Citizenship scale (Morais & Ogden, 2011).

Ewell and Jones (1991) point out that the differences between indirect and direct evidence are really differences in degree more than differences in kind. Ultimately, that continuum is based on the extent of perceived reliability and validity in tapping a construct of interest. Validity refers how accurately or "truthfully" a measure captures a concept (such as "social responsibility") (Carr, Boyle, Cornwell, Correll, Crosnoe, Freese, & Waters, 2021), or "the usefulness of information on whatever an assessment is intended to assess" (Suskie, 2018 p28). Reliability is concerned with how consistently a tool can measure a construct (e.g., Do different raters complete or interpret the same rubric consistently or similarly or do different test-takers interpret the items in the same way?) Indirect measures are considered less convincing because of the threats to reliability and validity somewhat intrinsic to their measurement–self-reported, unvalidated surveys or simply recording completion of an activity such as an internship.

In that sense, validated psychometric instruments, even though most are "surveys," arguably fall closer to direct evidence on the continuum. Indeed, the discipline of psychometrics is the science of

developing assessment instruments that attempt to connect observable phenomena (e.g., responses to an ambiguity tolerance scale) to theoretical attributes (e.g., “a capacity for nuance”) (Borsboom, n.d.). Or, if you prefer, “Psychometrics is the approximation of latent psychological processes by means of stochastic analysis at both the individual and population levels” (Molenaar, n.d.).

Establishing a validated psychometric instrument, scale or items follows a rigorous (and relatively standard) process, outlined well by DeVellis and Thorpe (2022): 1) Determine clearly what is to be measured, drawing on theory; 2) Generate an item pool, avoiding double-barreled, double negative, long or biased items, and response set problems, and aim for redundancy to facilitate later analyses and culling; 3) Establish the measurement and response scaling/categories and presentations; 4) Get expert review of the item pool, focusing on face validity, clarity, conciseness and adequate coverage of the construct in question.

Some researchers might also conduct “cognitive interviewing” to determine how potential respondents interpret the items (Step 5) and/or include validation items, such as a pre-existing social desirability scale to help detect undesirable responses (Step 6). After all of this, the first pilot is administered (Step 7) followed by statistical validation of the items, generally including item variance and distributions; reliability/internal consistency analyses and factor analyses for scales; tests for concurrent validity if using previously validated scales/items for testing (see Step 6); and maybe predictive validity. Items are then culled based on the analyses and final scales are established for replication and use (DeVellis & Thorpe, 2022).

Lumping this process in with “student and alumni surveys” is obviously inaccurate, yet no widely used model of direct and indirect evidence cites psychometric instruments within either category, implicitly, if not explicitly, considering them surveys, no different from a semester-end class survey/evaluation that asks simply, “Did this class improve your critical thinking?”

Acknowledging the rigor in developing and testing reliable and valid psychometric instruments opens the possibility that they do not fit in the current binary model of direct vs. indirect evidence, wherein “surveys” are placed squarely on the indirect side as a less convincing outcome (due in large part to legitimate methodological concerns of reliability and validity for many surveys and course evaluations). Psychometric instruments, however, are not typical examples of student or alumni surveys or end of semester course evaluations asking students to simply estimate what they learned or whether they achieved a certain SLO. It’s time to distinguish class evaluations and similarly structured student or alumni surveys from validated psychometric tools that utilize surveys to collect the evidence.

We must also remember that the construct (or SLO) of interest defines the most appropriate measure, not the other way around. It is the primacy of the SLO that guides the choice of measure best tapping into the construct of interest. When it is an attitude/value/disposition, properly identified psychometric instruments could very well be the most direct measure. Psychometric instruments by their definition should be both valid and reliable (certainly no less so than the majority of tests and rubrics currently considered direct), and thus can be considered direct evidence for that construct.

It is a worthwhile research question to ask whether the current direct/indirect dichotomy and its characterization of indirect evidence incentivizes instructors and assessors to instead work backward, first identifying direct evidence that can be feasibly collected and reported each semester, then moving to the SLOs that best fit that measure. Because it can be hard to feasibly assess attitudes/values/dispositions through current notions of direct evidence, they might be dropped for skills and knowledge tests more easily counted through tests and rubrics, which carry the coveted mantle of direct evidence.

The proposed 2x2 model presented below incorporates these concerns and suggestions, integrating the SLO type (Knowledge/Skills or Attitudes/Values/Dispositions) with the appropriate measure (Direct or Indirect). In doing so, it acknowledges there is a place for psychometrics to directly tap attitudes in the same way a test directly taps knowledge.

Table 2

The 2X2 Model of SLO Type x Evidence Type

	Direct	**Indirect**
Knowledge/ Abilities	-Scores on locally designed multiple-choice and essay tests/quizzes, accompanied by test blueprints describing what the tests assess -Score gains (“value-added”) -Rubrics that are specific and validated -Portfolios graded with rubrics -Observations of field work, internship performance, service learning, or clinical experiences -Capstone projects, senior theses, exhibits, or performances scored using a rubric -Pass rates or scores on licensure, certification, or subject area tests that assess key learning outcomes -Student publications and conference presentations -Employer and internship supervisor ratings of students' performance	-Internship completion -Service learning participation -Employment placement -Rubrics that are broad and unvalidated -Course grades and grade distributions -Percent of class time spent in active learning -Number of student hours spent on homework -Admission rates into graduate programs -Job placement rates and starting salaries -Alumni perceptions of career responsibilities and satisfaction -Student, alumni, and employer satisfaction with learning collected through surveys, exit interviews, or focus groups -Student ratings of their knowledge, skills, and reflections on what they

		have learned over the course of their program -Honors, awards
Attitudes/ Values/ Dispositions	-Validated psychometric scales and items -Student reflections on values and attitudes	-Course evaluations -Number of student hours spent on service learning -Unvalidated student and alumni Surveys -Number of student hours spent at intellectual or cultural activities related to the course

Table Note. Adapted from (Suskie, 2009, 2018; Palomba & Banta, 1999)

References

Borsboom, D. (n.d.) What is psychometrics? Retrieved from https://www.psychometricsociety.org/what-psychometrics May 2, 2024.

Carr, D., Boyle, E. H., Cornwell, B., Correll, S., Crosnoe, R., Freese, J., & Waters, M. C. (2021). The art and science of social research (2nd edition).W. W. Norton & Company.

DeVellis, R. F. & Thorpe, C. T. (2022). *Scale development: Theory and applications* (5th edition). SAGE.

Ewell, P. T. (1987). Establishing a campus-based assessment program. In Halpbern, Diane, F. (Ed.) *Student outcomes assessment: What institutions stand to gain*. CA: Jossey-Bass Inc.

Ewell, P. T. & Jones, D. P. (1991). Actions matter: The case for indirect measures in assessing higher education's progress on the National Education Goals. National Center for Education Statistics. October 7. Commissioned paper prepared for a workshop on Assessing Higher Order Thinking & Communication Skills in College Graduates (Washington, DC, November 17-19, 1991).

Furnham, A. & Ribchester, T. (1995). Tolerance of ambiguity: A review of the concept, its measurement and applications. *Current Psychology* 14, 179–199. https://doi.org/10.1007/BF02686907

Krislov, M. (2017). The Life-Shaping Power of Higher Education. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/views/2017/10/06/importance-liberal-arts-transforming-lives-essay. May 1, 2024.

Molenaar, P. (n.d.) What is psychometrics? Retrieved from https://www.psychometricsociety.org/what-psychometrics May 2, 2024.

Morais, D. B. & Ogden, A. C. (2011). Initial development and validation of the Global Citizenship Scale. *Journal of Studies in International Education, 15*(5), 445-466.

Palomba, C. & Banta, J. (1999). *Assessment essentials: Planning, implementing and improving assessment in higher education*. San Francisco, CA: Josey-Bass.

Paralkar, U. & Knutson, D. (2023). Coping with academic stress: Ambiguity and uncertainty tolerance in college students. *Journal of American College Health*, *71*(7), 2208–2216.

https://doi.org/10.1080/07448481.2021.1965148

Pargament, K., Feuille, M., Burdzy, D. The Brief RCOPE: Current Psychometric Status of a Short Measure of Religious Coping. *Religions*, *2*, 51-76. https://doi.org/10.3390/rel2010051

Rhodes, T. & Bergeron, J. (2017). Direct assessment methods. In *Enhancing assessment in higher education.* T. Cuming & M. David Miller (Eds). 49-77. Association for Institutional Research & Association of American Colleges and Universities: VA.

Suskie, L. (2009). Assessing student learning: A common sense guide (2nd edition). Jossey-Bass.

Suskie, L. (2018). Assessing student learning: A common sense guide (3rd edition). Jossey-Bass.

Swiecki, Z.i, Hassan K., Guanliang C., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gasevic, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3, 1-10

About the Author

Kevin Whiteacre, Director of Assessment, Office of Assessment, University of Indianapolis, whiteacrek@uindy.edu