Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2020, Vol. 39, No. 1, pp. 82–91

Deficiency, Contamination, and the Signal Processing Metaphor

Paul E. Newton, *Office of Qualifications and Examinations Regulation (Ofqual)*

Educational assessment involves eliciting, transmitting, and receiving information concerning the level of proficiency of a learner in a specified domain. With that in mind, it is perhaps surprising that the literature seems to make very little use of the signal processing metaphor. The present article begins by making a general case for greater use of this metaphor, as a simple and intuitive thinking tool for helping to explain how educational assessment works. The main body of the article extends this argument by demonstrating the utility of the metaphor in helping to explain how educational assessment can go wrong. During the 1980s and 1990s, Samuel Messick extensively discussed two major ways in which educational assessment can go wrong via construct-irrelevant variance and construct underrepresentation, respectively. Despite their importance, these concepts have had only a limited impact on the literature. Part of the reason for this may be a lack of clarity and comprehensiveness in their articulation. The present article aims to articulate these concepts more clearly and comprehensively, within a framework provided by the signal processing metaphor, reconfiguring them as signal contamination and signal deficiency, respectively.

Keywords: construct-irrelevant variance, construct underrepresentation, signal contamination/purity, signal deficiency/saturation, signal processing

As an international community of educational measurement professionals, we have become very proficient in modeling measurement outcomes. From Edgeworth (1888) to *Educational Measurement* (e.g., Haertel, 2006; Yen & Fitzpatrick, 2006), we have developed increasingly sophisticated psychometric approaches to modeling assessment results: classical test theory, generalizability theory, item response theory, and so on. This technology, if not science, of statistical modeling is now very well established. We have made a great deal of progress.

Unfortunately, the same cannot be said of our progress in modeling measurement processes. That Schmeiser and Welch (2006) opened their *Educational Measurement* chapter with the rhetorical question "Test development: art or science?" seems to reflect and to reinforce this impression. Rather than a robust technology of assessment design, we have a plethora of assessment traditions, which often seem to be remarkably resistant to progress.

Fortunately, things are beginning to change. The development of frameworks for evidence-centered design (ECD) epitomizes this direction of travel (e.g., Mislevy, 2007; Zieky, 2014). ECD frameworks provide language, concepts, and knowledge representations for designing, developing, and delivering educational assessments, based upon the organizing principle of evidentiary argument (Mislevy & Haertel, 2006). ECD incorporates powerful thinking tools for modeling the features and processes that comprise effective assessment procedures.

One of the biggest impediments to developing a robust technology of assessment design is the relative absence of models, or metaphors, for characterizing how educational assessments are supposed to work. It seems fair to say that, as a profession, we are still a few spanners short of a toolkit, in this respect. Furthermore, as I will explain shortly, some of our existing tools seem to be quite blunt, if not a little bent out of shape, which might help to explain why they are not deployed as frequently, or as successfully, as we might expect.

The purpose of the present article is to demonstrate the utility of a very simple thinking tool—the signal processing metaphor. This metaphor is not totally absent from the literature, although it is not widely utilized, which suggests that its true value remains to be appreciated.[1] I will attempt to demonstrate its value by using it to scaffold and, more importantly, to sharpen and straighten two of the most important, yet underutilized, concepts in the educational measurement toolkit—the ideas of construct-irrelevant variance (CIV) and construct underrepresentation (CUR).

How Educational Assessment Works

What is an educational assessment supposed to do? According to the Glossary of the *Standards for Educational and Psychological Testing*, an "assessment" is:

> Any systematic method of obtaining information, used to draw inferences about characteristics of people, [etc.]. (AERA, APA, & NCME, 2014, p. 216)

We can expand this definition by saying that the function of any educational assessment is to transmit information concerning the level of proficiency of a learner, from that learner

Paul E. Newton, Office of Qualifications and Examinations Regulation (Ofqual), Earlsdon Park, 53-55 Butts Road, Coventry, CV1 3BH, UK; paul.newton@ofqual.gov.uk

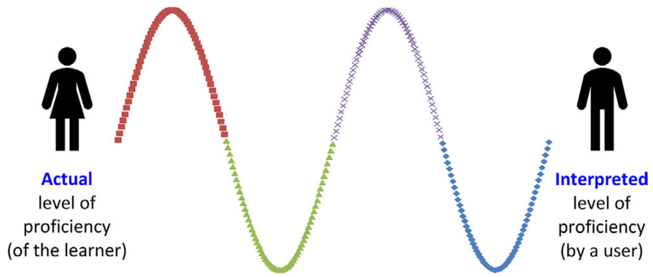

FIGURE 1. The signal processing metaphor. [Color figure can be viewed at wileyonlinelibrary.com]

to a user, that is, to someone who requires that information in order to make a decision.[2] An effective measuring technique, from this perspective, is one that ensures that the measurement information is emitted, transmitted, and received, as objectively as possible.

Figure 1 illustrates this process, thereby introducing the signal processing metaphor. It represents both a learner, whose proficiency is to be measured, and a user, who is ultimately responsible for the measurement interpretation.[3] Connecting the learner and the user is the signal, the function of which is to encode measurement information as objectively as possible. Notice that the signal takes on a different form from phase to phase. This is intentional. It is meant to convey that any measuring technique will encode measurement information via a succession of different formats, from the initial elicitation-performance phase to the final reception-interpretation phase. This is consistent with the image of educational assessment as a signal processing production line.

The idea of a production line, itself a metaphor, is based upon the observation that any assessment event can be subdivided into a number of logical steps, for instance:[4]

1. Elicitation: Multiple performances are elicited, from a learner, via assessment tasks, to provide a sample of evidence of proficiency.
2. Evaluation: Each performance in the sample is evaluated, by an assessor, in terms of what it implies about learner proficiency.
3. Combination: The set of performance evaluations is combined, by an aggregator, into a measurement result.
4. Interpretation: The measurement result is interpreted, by a user, to whom it has ultimately been delivered.

Each of these steps can therefore be associated with a principal output:

1. Elicitation output: a set of task performances for each learner—their performance profile.
2. Evaluation output: a set of evaluations for each learner—their evaluation profile.
3. Combination output: a result for each learner—their measurement result.
4. Interpretation output: an interpretation of the result—the measurement interpretation.

The production line metaphor indicates that an output from one step becomes the input to the next: a learner produces a profile of performances, which they metaphorically pass down the line to an assessor; the assessor produces a profile of evaluations, which they metaphorically pass down the line to an aggregator; and so on. Together, the two metaphors illustrate how any assessment event involves *presenting*, *extracting*, *interpreting*, and *re-presenting* measurement information (information concerning the proficiency of a learner) *multiple times* on the way to a final measurement interpretation.

We can think of each of these outputs as a new expression of the signal, as it travels from the learner to the user. During each phase, measurement information is encoded in the output (the signal) in a different format. In other words, the output from each step is a new, and differently formatted, representation of the learner's level of proficiency: from the learner's own representation of their own level of proficiency—represented in their performance profile; to the user's representation of the learner's level of proficiency—represented in the final measurement interpretation. If each successive representation is faithful to the one that preceded it, then the measurement information will be transmitted objectively, and the final measurement interpretation will be accurate.

Accuracy is one thing, but usefulness is essential too. This is ensured by a prior step, which occurs at the outset of the design process, during which the proficiency that needs to be measured—the target proficiency—is specified (see Newton, 2017a, 2017b):

0. Clarification: Measurement objectives are clarified. In particular, the proficiency that needs to be measured—the target proficiency—is articulated. Principal output = a proficiency specification.

The proficiency specification is the cornerstone of any assessment procedure. Once articulated, it becomes the point of reference for each of the steps that follows. Whenever information is presented, extracted, interpreted, and represented during the operation of a measuring technique, it is essential that each step preserves representational faithfulness *simultaneously*: to the proficiency of the learner (the "thing" that is being measured); and to the proficiency specification (the construct through which the measurement is given meaning).[5] If so, then representational faithfulness will endow the assessment procedure/measuring technique with both objectivity *and* intersubjectivity: objectivity—to the extent that measurement information can always be traced back to the proficiency of the learner, and only to the proficiency of that learner; and intersubjectivity—to the extent that the measurement interpretation can always be traced back to a common (set of) measurement standard(s), represented as part of the proficiency specification (see also Perie & Huff, 2016). Luca Mari has explained how these two qualities, objectivity and intersubjectivity, represent defining characteristics for measurement, within the metadiscipline of metrology (e.g., Mari, 2000, 2013).

There is a risk, when employing the signal processing metaphor, that its language may create an impression of assessment as a passive process; whereby information is "emitted" from an isolated learner, and then simply "travels" to an isolated user. In fact, the intention behind the metaphor is to support exactly the opposite conclusion, that is, that assessment is an inherently constructive process; indeed, a structured, co-constructive process. So, it is important that this message is emphasized. Note that, during step 1, the learner is the central agent, and it is ultimately up to them to construct a faithful representation of their own level of proficiency, facilitated by an assessment developer, who has helped to create the conditions (including the instrument) through which they are empowered to construct that faithful representation. It is also worth mentioning that the learner will become one of the multiple users of their own assessment result, once it has been delivered (other users might include teachers, parents, administrators, counselors, coaches, and so on). So, the learner will adopt an active role in both presenting *and* interpreting measurement information, concerning their own level of proficiency. Again, all of these considerations need to be factored into the design of the assessment procedure, broadly conceived; to ensure the accuracy of the coconstruction of measurement interpretations for subsequent uses.

What Could Possibly Go Wrong?

For at least three decades, the educational measurement literature has recognized two key thinking tools for identifying alternative ways in which educational assessment can go wrong. Messick described these twin concepts as follows:

> The two major threats to construct validity are *construct underrepresentation* – that is, the test is too narrow and fails to include important dimensions or facets of the construct – and *construct-irrelevant variance* – that is, the test contains excess reliable variance, making items or tasks easier or harder for some respondents in a manner irrelevant to the interpreted construct. (Messick, 1989a, p. 7)

Despite their fundamental importance, the status of these concepts in the pantheon of the psychometrician (to borrow a metaphor from Ebel) does not yet appear to have been universally recognized. In the following subsections, I will explain what I mean by this, and why I think that this might be so.

The Importance of These Concepts

The idea that there are two major threats—CUR and CIV—was identified by Messick in the third edition of *Educational Measurement* (Messick, 1989b), although he actually attributed the idea to Cook and Campbell (1979). Messick used these concepts repeatedly both in his chapter and in his subsequent papers (e.g., Messick, 1994, 1995a, 1995b, 1996, 1998). The *Standards* now locates these concepts at the heart of its introduction to validity (AERA et al., 2014, pp. 12–13).

As validity is generally considered to be the most fundamental consideration in designing, developing, delivering, and evaluating educational assessments (AERA et al., 2014, p. 11), we might assume that the twin threats ought, by now, to have acquired a high profile in the educational measurement literature, as core organizing principles, perhaps. Yet, while they are not unfamiliar concepts, at least among measurement professionals, they do not seem to be especially well used. Indeed, although they *are* occasionally used as organizing principles, this seems to be more the exception than the rule (see McCallin, 2006, for a valuable exception). Interestingly, although the editor of the fourth edition of *Educational Measurement* highlighted Messick's prior discussion of the twin threats (Brennan, 2006, p. 2), these concepts featured less prominently in Kane's chapter, *Validation*, in the fourth edition, than in Messick's corresponding chapter, *Validity*, in the third (see Kane, 2006; cf. Section 3.4 on *Trait Underrepresentation and Irrelevant Variance*). Indeed, according to its index, the only chapter of *Educational Measurement* to make more than passing reference to CIV-concerned performance assessments was Lane and Stone (2006).

Although textbook index entries can provide only limited evidence of how well-utilized such concepts are in the literature, it is still worth mentioning that the index of *Modern Measurement* (Osterlind, 2010) contains no reference to either CUR or CIV. Neither does the index of *Educational Assessment of Students* (Brookhart & Nitko, 2019), nor *Educational Testing & Measurement* (Kubiszyn & Borich, 2015). The terms do not appear in the index of *Measurement and Assessment in Teaching* (Miller, Linn, & Gronlund, 2013), although they are discussed in Chapter 4. The terms CUR and CIV do appear in the index of *Measurement and Evaluation in Psychology and Education* (R. M. Thorndike & Thorndike-Christ, 2010), directing the reader to page 185, where they are explained. Likewise for *Measurement and Assessment in Education* (Reynolds, Livingston, & Wilson, 2010, p. 124), and for *Educational Assessment* (Hogan, 2007, p. 68). Having said that, none of these key textbooks makes extensive use of either CIV or CUR.

If Messick is right, then the two major threats ought to be fundamental to any discussion of assessment design, development, delivery, or evaluation. Yet, their use in the educational measurement literature does not seem to reflect this. This prompts me to consider why. One reason, I suggest, is that the concepts are simply not user-friendly enough. More specifically, the manner in which they tend to be articulated is confusing, lacking both clarity and completeness. In response to this challenge, I will attempt to reconfigure the twin threats, within a framework provided by the signal processing metaphor, in a manner that renders them clearer, more comprehensive, less confusing and more useful.

Problems With These Concepts

The most immediate problem with the twin threats is how confusingly they are labeled. For instance, what exactly is it that is underrepresented in CUR? And what exactly is it that varies irrelevantly in CIV? Although, in both cases, the idea of a construct is clearly implicated, the term "construct" is used in all sorts of different ways in the literature (Slaney & Garcia, 2015), and this inevitably renders any compound term that incorporates it at least somewhat ambiguous. Having said that, Messick clearly relates the term "construct" to the meaning, or the interpretation, of test scores. Presumably, then, CUR implies that test score meaning is somehow underrepresented, while CIV implies that test score meaning is somehow subject to irrelevance. But how so?

The quotation from Messick (1989a, p. 7, see above) specifies CUR in terms of "the test" being too narrow, and specifies CIV in terms of "the test" containing excess reliable variance. With respect to CUR, he seems to suggest that it occurs when the set of tasks that comprise the test fails to represent the full breadth of its construct. If so, then it will not necessarily be legitimate to interpret test score meaning in terms of the (full) construct.

With respect to CIV, though, it is less clear what Messick is suggesting. The idea that the *test* somehow contains excess variance "making items or tasks easier or harder for some respondents in a manner irrelevant to the interpreted construct" is grammatically challenging. This is because the concept of "variance" is associated with test scores, and test scores have no impact on how respondents perform test tasks. In short, what Messick seems to want to describe, using the

CIV label, is a *cause* of invalidity that mediates candidates' interactions with test tasks. Yet, what he seems to refer to, with the CIV label, is a *consequence*: irrelevant variance in test scores. The idea that Messick was primarily interested in describing a cause of invalidity is also supported by his claim that there are "two basic kinds of construct-irrelevant test variance" (Messick, 1989b, p. 34)—construct-irrelevant difficulty and construct-irrelevant easiness—which he specifically developed in terms of construct-irrelevant *sources* of difficulty and easiness, related to task demands and features, as well as to test-taking strategies.

Having said that, there does seem to be something important (to Messick) about the way in which he frames CIV in terms of irrelevant *score* variance; more specifically, in terms of irrelevant *reliable* score variance. In doing so, he seems to be leaving conceptual space for the concept of reliability *alongside* the twin threats; as though reliability, CUR, and CIV were mutually exclusive and (perhaps even) collectively exhaustive concepts, within a broad construct validity framework. For example,

> In evaluating the adequacy of construct measurement, two closely connected points need to be disentangled. One is that tests are not only imprecise or fallible by virtue of random errors of measurement [REL] but also inevitably imperfect as exemplars of the construct they are purported to assess. Tests are imperfect measures of constructs because they either leave out something [CUR] that should be included according to the construct theory or else include something [CIV] that should be left out, or both. (Messick, 1989b, p. 34; square brackets added)

Haladyna and Downing (2004) extended this idea:

> CIV is error variance that arises from systematic error. A good way to think about systematic error is to compare it with random error. If we were to write the linear model representing what we know about random and systematic errors, the model would be:
>
> $y = t + e_r + e_s$,
>
> where y is the observed test score for any student, t is the true score, e_r is random error, and e_s is systematic error due to CIV. (p. 18)

In fact, even from a purely statistical perspective, this is only partially correct, because e_s ought to be subdivided into a component for CIV and a component for CUR, as both are potential sources of bias. Incidentally, this raises the question of why Messick chose not to refer to CUR as "CUR (systematic error) variance" (or suchlike) to parallel "CIV (systematic error)." To be fair, CIV and CUR are unwieldy enough labels as they are! Yet, the disparity in their labeling is surely not insignificant.[6]

Two Perspectives on These Concepts

As presented by Messick, the twin threats can be understood both statistically and causally. Statistically, they capture facets of bias: CUR (systematic error) variance and CIV (systematic error).[7] Causally, they indicate the explanation of those facets. Notice that something about the idea of CUR seems to steer us more toward a causal perspective, while something about the idea of CIV (which must include the use of "variance" in its label) seems to steer us more toward a statistical perspective. This observation is nicely illustrated in the definitions of these concepts in the Glossary of the *Standards* (AERA et al., 2014, p. 217):

> The extent to which a *test* fails to capture important aspects of the construct domain that the test is intended to measure, resulting in test scores that do not fully represent the construct. [CUR – emphasis added]
>
> "*Variance* in test-taker scores that is attributable to extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation." [CIV – emphasis added]

Perhaps, then, we should not be surprised if people find these concepts confusing . . . they literally do not *sound* as though they complement each other as neatly as they ought to.

Now, although Messick's characterization suggests that CUR and CIV can be understood from both perspectives, their *primary* function (from both perspectives) is to indicate a particular causal explanation of how the assessment procedure has gone wrong. In other words, the *primary* perspective on the use of these terms is causal. This leads me to question whether it might actually be clearer to define both of these concepts *purely* at the causal level such that neither should be *defined* in terms of their implications for variance, whether systematic or unsystematic.

In fact, if we were to do so, the payoff in terms of clarity would be substantial. This is because the concepts of irrelevance and underrepresentation (viewed from a causal perspective) are entirely independent of the concepts of systematic error variance and unsystematic error variance (viewed from a statistical perspective). In other words, *both* systematic *and* unsystematic error variance (i.e., both bias and unreliability) can be attributed to *either* underrepresentation *or* irrelevance. Furthermore, whether or not the error variance arising from any particular cause is properly classified as either systematic or unsystematic has nothing to do with the nature of that cause. Instead, this is simply a matter of whether the cause is presumed to operate consistently or inconsistently across occasions, raters, group members, and so on. Bearing all this in mind, the idea that we should *only* apply the idea of "construct-irrelevant corruption" when it appears to operate systematically—thereby resulting in systematic error variance—would seem to be unduly (if not bizarrely) restrictive, especially for a concept whose primary function is to support causal explanation.

A Subtle Reconfiguration

Interestingly, the path towards a subtle reconfiguration was already trodden by Messick in his seminal chapter. Yet, the paving stones in this path were laid by Brogden and Taylor (1950), rather than Cook and Campbell (1979).[8] Brogden and Taylor were actually describing the evaluation of criterion measures, and in a section entitled *Criterion Bias*, they observed that

> Imperfections or bias in the criteria may be classified as:
>
> (1) *Criterion Deficiency* – omission of pertinent elements from the criterion.
>
> (2) *Criterion Contamination* – introducing extraneous elements into the criterion.
>
> (3) *Criterion Scale Unit Bias* – inequality of scale units in the criterion.

(4) *Criterion Distortion* – improper weighting in combining criterion elements." (Brogden & Taylor, 1950, p. 161)

Messick noted the direct parallels between criterion deficiency and CUR, and criterion contamination and CIV, respectively (e.g., Messick, 1989a, p. 70). Indeed, the *Standards* now puts "construct deficiency" in parenthesis after CUR, and "construct contamination" in parentheses after CIV (AERA et al., 2014, p. 12).

These two core concepts—deficiency and contamination—can be made to work in close partnership, at a causal level, by applying the signal processing metaphor. As noted earlier, what needs to be emitted, transmitted, and received via a measuring technique is measurement information, concerning the proficiency of a learner. Thus, the two core concepts are analogous to measurement information deficit, and measurement noise, meaning that the twin threats can be reconfigured as *signal deficiency* and *signal contamination*, respectively. These terms are very much more specific than their "construct" counterparts. In both cases, it is clear that it is the signal that is deficient and the signal that is contaminated. Deficiency means that the signal represents some, but not all, of the measurement information related to the target proficiency. Contamination means that the signal is influenced by something other than the target proficiency.

From Quality Threats to Quality Criteria

A desire to apply these concepts in the contexts of assessment design, development, and delivery leads me to transform them into positive quality criteria. Rather than seeking to avoid signal deficiency and signal contamination, our twin aspirations are to maximize *saturation*—such that the signal tells the whole truth about learner proficiency; and *purity*—such that the signal tells nothing but the truth about learner proficiency. These criteria refer to the quality of the signal—how representationally faithful it is to the measurement information that needs to be conveyed—as each assessment event cycles through multiple iterations of presentation, extraction, interpretation, and representation, on the way to a final measurement interpretation.

Signal Saturation

The idea of transmitting *all* of the relevant information has particular connotations in educational measurement; in particular, because our target proficiencies (the "things" that we need to measure) are composite attributes (see Maul, 2013). The fact that each target proficiency will be defined in terms of distinguishable elements (of knowledge, skill, etc.) means that educational measurement is likely always to involve some degree of both analysis (decomposition) and synthesis (recomposition). Bearing this in mind, signal saturation can be subdivided into at least four distinguishable subcriteria:

1. Completeness: The signal represents the full breadth of the target proficiency (i.e., conveys information on all elements).
2. Balance: The signal proportionately represents the elements of the target proficiency (i.e., conveys information in a manner that preserves proportionality).
3. Interconnectedness: The signal represents how the elements of the target proficiency are structured and integrated (i.e., conveys information on the organization and coordination of elements).
4. Detail: The signal represents the target proficiency in full resolution (i.e., conveys information in a manner that preserves trueness-to-life).

Note how these four quality criteria contain two that resonate strongly with traditional criteria for developing tests (completeness and balance), alongside two that resonate strongly with more recently proposed criteria for developing performance assessments (interconnectedness and detail).

The interconnectedness criterion reflects an idea that has become widely recognized as a point of principle, but that has not yet become deeply embedded in practice: that educational assessments ought (somehow) to recognize levels of proficiency as locations along trajectories, on a path from relative novice to relative expert (Glaser, 1981; Perie & Huff, 2016). This is not simply a matter of representing the amount of knowledge acquired by a learner; as though learning was simply a matter of accumulating more and more discrete elements. What also needs to be represented is the complexity of the organization and coordination of that knowledge (see Bereiter & Scardamalia, 2005; Eraut, 2004; Messick, 1984; Mislevy, 1993; Sadler, 1987; van der Vleuten, 1996; Wood & Power, 1987). So, the idea of interconnectedness refers to the aspiration of conveying measurement information concerning the structure, organization, integration, and coordination of knowledge and skill within a domain.[9]

The detail criterion refers to the aspiration of conveying measurement information in a manner that is true-to-life, as opposed to being artificial, in the sense of being abstracted-away-from-life. The greater the degree of abstraction, the lower the resolution of the measurement information, and the more deficient the signal becomes. This resonates with the literature on authenticity in educational assessment (e.g., Spolsky, 1985; Wiggins, 1989). However, the detail criterion is not a matter of the assessment context failing to "look like" real life, in some woolly, superficial sense (see Cumming & Maxwell, 1999). It is a matter of an assessment event failing to represent the target proficiency in its full subtlety and richness.

Signal Purity

Complementing signal saturation, signal purity describes the absence of construct-irrelevant corruption. In the revised framework, it is conceptualized in terms of measurement information, rather than scores, which is carried via a signal; and it refers to the absence of contamination, that is, to the absence of noise, rather than to irrelevant score variance. In addition to maximizing signal saturation, effective assessment design aims to maximize signal purity; that is, to ensure that the various processes of presenting, extracting, interpreting, and representing measurement information—which operate during any assessment event—are engineered so as to minimize contamination by influences beyond the target proficiency.

Unlike signal saturation, signal purity cannot be quite so neatly subdivided, although scholars have proposed a number of alternative schemes over the years. Messick, for instance, subdivided CIV into construct-irrelevant sources of difficulty and construct-irrelevant sources of easiness (Messick, 1989b, p. 34). Arguably, a better way of describing this would be to distinguish between outcome-deflating contaminants and outcome-inflating ones. This is because there is more to CIV than difficulty or easiness, interpreted in terms of the demands associated with assessment tasks, either relevant or irrelevant. For example, even after learners have completed all of their assessment tasks, raters can still introduce outcome-inflating/deflating contamination, as can those who interpret assessment results.

Unfortunately, once unpacked, the distinction that Messick foregrounds for CIV seems to be fairly arbitrary. For instance, a contaminant that is outcome-deflating for one assessee might be outcome-inflating for another (e.g., the choice of soccer/football as the context for a math problem); so, it is unclear how the distinction is conceptually important. More importantly, the fact that any instance of CUR (i.e., signal deficiency), could equally be characterized as either outcome-inflating or outcome-deflating underscores the triviality of this distinction.

Haladyna and Downing recommended an alternative approach to unpacking CIV, "providing a simple taxonomy for classifying variables that produce" it (Haladyna & Downing, 2004, p. 19):

1. uniformity and types of test preparation (e.g., extensiveness of test preparation),
2a. test development (e.g., poor item quality),
2b. test administration (e.g., adherence to time limits), and
2c. test scoring (e.g., failure to equate test forms),
3. students (e.g., fatigue); and
4. cheating (e.g., hiring professional test takers as substitutes).[10]

Although this taxonomy is helpful in identifying signal contamination threats presented by a variety of factors, it is important to note that it is not of the same kind as the scheme proposed earlier, for signal saturation, which was based upon an analysis of the different ways in which a signal might be deficient of measurement information. In fact, it is unclear whether the same kind of scheme could be provided for signal purity, and there are no obvious clues within the literature as to how such a scheme might be constructed. This may be because saturation relates to the inclusion of genuine measurement information, the nature of which is relatively bounded and specifiable; whereas purity relates to the exclusion of anything that is not genuine measurement information, which is unbounded and unspecifiable.

Although the Haladyna and Downing taxonomy is helpful, it is far from complete. For instance, referring back to the production line metaphor presented earlier, it does not deal adequately with threats arising during either: step 0 (clarification—the target proficiency is faithfully represented by the proficiency specification); or step 4 (interpretation—the measurement result is faithfully represented by the measurement interpretation). The production line metaphor therefore helps to expand our analysis of potential signal contaminants. Perhaps, the most frequently overlooked source of signal contamination is the misinterpretation of a measurement result by a result user (e.g., a selector in a human resources department reading results presented on a resumé). Assessment procedures should be designed to be as contamination-resistant during step 4 (interpretation) as they are during all of the other steps. In particular, attention needs to be paid to the "form-giving role of values in determining or distorting the meaning of score interpretations per se" (Messick, 1989b, p. 59). For instance, the labels used to describe the target proficiency, and the labels used to describe each of the standards associated with the proficiency scale, all need to be articulated so as to maximize the likelihood of signal saturation and signal purity, when results are interpreted by users.

Validity

Although this section has been framed in terms of quality threats and quality criteria—referring to the quality of the signal at the heart of each assessment event—it is helpful to be clear that this is quality-as-validity, as Messick understood the concept of validity; and that signal deficiency and signal contamination are his two major threats to validity. Of course, using the term *validity* can be problematic, because even scholars of validity disagree radically over the best way to use this word (see Newton & Baird, 2016). Having said that, deficiency and contamination would probably count as the two most important threats to validity under almost any of the competing definitions.

How These Thinking Tools Help Us

How does the signal processing metaphor, alongside the reconfigured concepts of signal deficiency and signal contamination, help to advance the theory and practice of educational measurement? After all, they are pretty simple thinking tools, and they are not especially original. The following subsections conclude the present article by summarizing how these thinking tools can help us, specifically, in terms of clarity, comprehensiveness, and generativity.

They Facilitate Clear Thinking

What the signal processing metaphor provides us with is a single, coherent framework within which to locate Messick's two major threats. By adopting exactly the same perspective on these concepts, they are rendered more consistent and thereby more comprehensible. According to this reconfiguration:

Signal deficiency occurs when a measurement signal—which conveys measurement information, concerning the proficiency of a learner, from that learner to a user via a measuring technique—becomes deficient of measurement information; that is, when an aspect of the proficiency that needs to be measured is underrepresented in the signal.

Signal contamination occurs when a measurement signal—which conveys measurement information, concerning the proficiency of a learner, from that learner to a user via a measuring technique—becomes contaminated by noise; that is, when the signal represents factors other than the proficiency that needs to be measured.

Defining the two major threats thus, within a signal processing metaphor, helps to clarify how perfectly they complement each other. Notice that they are both defined at the level of providing a causal explanation for the success or failure of an individual measurement. Analysis at this causal level is clearly separated from analysis at the statistical level, reflecting the fact that analysis at the causal level (where the focus is upon the nature of error for individual measurements) is logically distinct from—and typically prior to—analysis at the statistical level (where the focus is upon the prevalence and consistency of errors across measurements).

These thinking tools also help to clarify that causal explanations do not have to be framed in terms of either underrepresentation or irrelevance, since both will often be implicated. Thus, where Haladyna and Downing (2004) identified cheating as a cause of CIV, it is worth noting that cheating will typically implicate both deficiency and contamination. For instance, when an individual impersonates a candidate, unlawfully taking a test on their behalf, the diagnosis of the error implies a total lack of measurement information (on the candidate), such that the observed signal comprises nothing but noise.

Finally, these thinking tools also help to facilitate clarity by recognizing that, during any assessment event, the measurement signal cycles through multiple phases of information presentation, extraction, interpretation, and representation. This model emphasizes how deficiency and contamination are *ever-present* threats, right from the outset of the design process; and through to each and every time that an individual measurement result is interpreted. This also helps to explain why they facilitate more comprehensive thinking.

They Facilitate Comprehensive Thinking

The ways in which CUR and CIV have traditionally been articulated have tended to encourage an unduly narrow focus. The focus of CUR has tended to be upon a poorly constructed test, whereas the focus of CIV has tended to be upon (the potential cause of) undesirable score variance. The concepts of signal deficiency and signal contamination widen that focus substantially, highlighting threats that arise during each of the phases through which a measurement signal cycles.

The most obvious failure of both CUR and CIV is that neither encourages the assessment designer to look beyond the measurement result itself. Yet, some of the most egregious causes of error occur during the final step in the assessment cycle, during which a measurement result is converted into a measurement interpretation, by a user with a decision to make. Clearly, a user might interpret the result to mean just one aspect of the target proficiency (deficiency), or they might interpret the result to mean something other than the target proficiency (contamination). So, the twin threats have a critical role to play in reminding assessment designers to establish controls that maximize the likelihood of full interpretation, and minimize the likelihood of misinterpretation.

The concept of signal deficiency also facilitates more comprehensive thinking about quality criteria for educational assessments, particularly through the analysis of alternative ways in which a measurement signal might be deficient—via incompleteness, imbalance, disconnectedness, or abstraction. This analysis implies that completeness, balance, interconnectedness, and detail are *general* quality criteria, against which *any* educational assessment ought to be judged, regardless of its format (Messick, 1994; cf. Baartman, Bastiaens, Kirschner, & van der Vleuten, 2007; Linn, Baker, & Dunbar, 1991). This brings the focus of the assessment designer back to where it ought to be—upon the best possible set of trade-offs for the situation in hand. A multiple-choice test might well be judged negatively against the criteria of interconnectedness and detail, while simultaneously being judged positively against the criteria of completeness and balance. But what ultimately needs to be judged is whether this is the optimal trade-off to make, for the situation in hand, which will also need to take into account its susceptibility to contamination by noise, as well as pragmatic considerations such as cost.

The introduction of response process evidence to the validity chapter of the *Standards* (AERA, APA, & NCME, 1999) has helped us to think more comprehensively about validation. Rather than restricting our evaluation efforts to the nature of our assessment instruments (e.g., content evidence), the *Standards* now emphasizes how we should also investigate how assessees respond to those instruments (e.g., process evidence). However, the process by which an assessee responds to an assessment task is just one of multiple information processing steps that occur during any assessment event. As the production line metaphor emphasizes, each of these steps is a worthy candidate for investigation, as part of a comprehensive validation program (Newton, 2019). Thus, our signal processing thinking tools are consistent with a far broader perspective on validation than even the current edition of the *Standards* recommends (see also Newton, 2016).

They Facilitate Generative Thinking

With the increasing automation of assessment processes—item development, test administration, response scoring, and so on—it is easy to forget that assessment is fundamentally an interpretative, meaning-making exercise, which enables one human (the user) to appraise another (the learner). The signal processing metaphor helps to redress this balance, by describing the entire enterprise of assessment in terms of a succession of information processing challenges. This helps us to think a little differently about educational measurement, and productively so.

One useful insight is that the threats of deficiency and contamination may be of a different order as well as of a different kind. Specifically, the threat of low measurement information is potentially more challenging than the threat of high noise, assuming that we can effectively hone our information-extraction mechanisms. This point is primarily a logical one: if our measurement information is extremely impoverished, then we are unlikely to be able to deliver an accurate measurement result, even in the complete absence of noise. Yet, if we have measurement information in abundance (i.e., our signal is saturated with it), then we have a fighting chance of being able to recover it, even when faced with a large amount of noise.

This is a helpful way of thinking about the difference between standardized testing and performance assessment. Standardized tests aim to eliminate as much noise as possible during step 1 (elicitation), but this also means losing a considerable amount of genuine measurement information (especially concerning interconnectedness and detail). Batteries of complex performance assessments, on the other hand, aim to capture as much genuine measurement information as possible during step 1 (elicitation), while necessarily also accumulating a substantial amount of noise. The accumulation of noise is not necessarily an insurmountable problem; just as long as the measuring technique has mechanisms in place through which to recover the information from a noisy signal. These are exactly the kinds of mechanisms that medical education assessment professionals are currently grappling with, to enable assessors to better understand, recognize, and manage their judgmental biases during oral and practice-based assessment (e.g., Schmidt & Schuwirth, 2013; Schmidt, Schuwirth, O'Keefe, & King, 2016).

The concepts of saturation (cf. deficiency) and purity (cf. contamination) have been introduced to explain how representational faithfulness can be maintained, across successive phases of an assessment event, to ensure that the final measurement interpretation will be accurate. During each step, the measuring technique needs to ensure that successive outputs—performances, evaluations, result, and interpretation—remain faithful to the learner's actual level of proficiency. So, this begs the question of what it might mean for each successive representation to be faithful, which opens up a variety of interesting questions for the assessment designer.

It is perhaps easiest to think about representational faithfulness in terms of resemblance, and this *is* an importance criterion, particularly in relation to step 1 (elicitation). In other words, it seems reasonable to say the more closely the learner's performance resembles the full breadth of their proficiency, and nothing but their proficiency, the better, because it maximizes the amount of genuine measurement information available. Resemblance also seems to be important in relation to step 4 (interpretation). In other words, the user needs to be able to reconstruct an image of the learner, on the basis of their result, and this reconstruction needs to resemble the target proficiency as closely as possible.

Yet, representational faithfulness is not always a matter of resemblance. For instance, there is no sense in which the measurement result—a Grade B, say—actually *resembles* the proficiency of the learner. It is just a symbol, of course. So, representational faithfulness means something different during this phase of an assessment event. More interesting still, faithful representation sometimes explicitly departs from the resemblance "ideal." In his discussion of scientific representation, van Fraassen (2008) paid considerable attention to the value of (selective) distortion, occlusion, abstraction, and suchlike. The success of a representation, he argued, sometimes relies upon intentional departure from resemblance, as occurs when a sculptor of large sculptures renders upper parts unnaturally large and lower parts unnaturally small. In doing so, the sculptor intentionally introduces imbalance (in my lexicon) to the representation, specifically in order to ensure that any onlooker (i.e., "up-looker") will receive the visually presented information successfully. What this suggests is that CUR-deficiency and CIV-contamination, which have been conceptualized within the measurement literature *purely* as threats, may not necessarily *always* be so. Considering the potential of selective deficiency and selective contamination to enhance representational faithfulness may begin to open up new and interesting avenues of research and analysis.

Acknowledgments

I would like to thank Dylan Wiliam for his insightful feedback on an early draft of this article. I also thank to the *EM:IP* reviewers for their constructive comments.

Notes

[1] Cronbach's discussion of fidelity versus bandwidth (Cronbach, 1989) and Wiliam's discussion of fidelity versus disclosure (Wiliam, 1992) represent interesting exceptions; although neither Wiliam nor Cronbach developed these ideas significantly.

[2] For the purpose of explanatory simplicity, the present article focuses solely upon the assessment of proficiency—typically referred to in the North American literature as summative achievement testing—which aims to represent a learner's level of proficiency in a domain of learning via an assessment result.

[3] Although the idea of an "actual" level of proficiency begs more philosophical questions than could be addressed within the present article, we might think of it simply as a useful idealization, which is not committed to any particular modeling strategy (e.g., generalizability theory, item response theory).

[4] It is worth mentioning two features of this particular characterization. First, the intention is to characterize the essential functional logic that underpins educational assessment, generally, rather than to stress the importance of particular agents or mechanisms. For instance, within a standardized testing context, steps 2 and 3 might be automated via optical character recognition and computing technology, and the combination step might need to be augmented by a mechanism for linking standards. Alternatively, within a performance assessment context, a single person might take the role of both assessor and aggregator, merging these roles via holistic judgments of proficiency level. Second, the intention is to focus upon what the assessment itself needs to do, rather than upon the role played by key facilitators (e.g., specification developers, item writers, test compilers, psychometricians). The focus is therefore upon what these facilitators need to facilitate (through specification development, item writing, test compilation, and so on). This helps us to dig below the surface of alternative assessment traditions.

[5] It is helpful to draw a subtle distinction between an assessment procedure and a measuring technique. An assessment procedure is the set of specifications that governs the entire activity of measuring. A measuring technique is the mechanism through which learners are measured, which comprises a set of apparatus and processes. In terms of the educational assessment life cycle—design, develop, deliver, and evaluate—the assessment procedure is produced during the design stage, whereas the measuring technique is produced during the development stage. Subsequently, during the delivery stage, the measuring technique is operated, to deliver measurements, that is, measurement interpretations.

[6] It is tempting to associate a lack of measurement information with a lack of variance. If so, then, perhaps, this is why Messick did not explicitly link the concept of CUR to the concept of variance—whether systematic or unsystematic—because an absence of information renders talk about true scores, error scores, and variance, redundant. In fact, though, this is not quite right. This is because an assessment event will always produce a signal; and it will always be possible to describe the quality of that signal (once a result has been produced) in terms of truth, error, and variance. If, for instance, a high-achieving examinee decided to flunk a test, by giving up after the second item, then we could understand this in terms of an absence of measurement information. Yet, despite there being very little information available, the assessment process would still construct a signal (a highly deficient signal), and it would still deliver a result (a highly inaccurate result).

[7] I am using the term *bias* in the sense of the second of the two definitions of this term within the Glossary of the Standards: "2. In statistics or measurement, systematic error in a test score" (AERA et al., 2014, p. 216). (The first sense refers explicitly to CIV and CUR, incidentally.)

[8] In fact, the idea underlying both CIV and CUR can be traced back a very long way. As Thorndike said of intelligence tests: "Our tests are also diluted or contaminated by the influence of other things than intelligence, such, perhaps, as speed of writing or effort to do one's best, or excitement and worry caused by the examination, or special circumstances of training. The ideal should be to sample all of interest [cf. CUR] and nothing but intellect [cf. CIV]" (E. L. Thorndike, 1925, pp. 6–7; square brackets added).

[9] In recent years, test designers and evaluators have attempted to do greater justice to the idea of interconnectedness by transforming one-dimensional sampling frameworks into two-dimensional ones (e.g., Content domain × Cognitive domain). A common approach is to deconstruct the cognitive domain hierarchically (e.g., via Bloom's Taxonomy), assuming that each content area ought to be assessed at each level of cognitive complexity (e.g., knowledge/comprehension, application/analysis, and synthesis/evaluation). This amounts to attempting to address interconnectedness via completeness and balance. Yet, the very idea of domain deconstruction is inherently fragmentary, and may end up completely neglecting, or negating, critical information concerning structure, organization, integration, and coordination across the elements of the target proficiency.

[10] Incidentally, their taxonomy is quite similar to one suggested 50 years earlier by R. L. Thorndike (1951, p. 568) in his chapter on reliability.

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing* (5th ed.). Washington, DC: American Educational Research Association.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing* (6th ed.). Washington, DC: American Educational Research Association.

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review*, *2*(2), 114–129.

Bereiter, C., & Scardamalia, M. (2005). Beyond Bloom's taxonomy: Rethinking knowledge for the knowledge age. In M. Fullan (Ed.), *International handbook of educational change: Fundamental change* (pp. 5–22). Dordrecht, The Netherlands: Springer.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed). *Educational measurement* (4th ed., pp. 3–16). Washington, DC: American Council on Education/Praeger.

Brogden, H. E., & Taylor, E. K. (1950). The theory and classification of criterion bias. *Educational and Psychological Measurement*, *10*(2), 159–183.

Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students* (8th ed.). Upper Saddle River, NJ: Pearson.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cronbach, L. J. (1989). Lee J. Cronbach. In G. Lindzey (Ed.). *A history of psychology in autobiography* (Vol. 3, pp. 63–93). Stanford, CA: Stanford University Press.

Cumming, J. J., & Maxwell, G. S. (1999). Contextualising authentic assessment. *Assessment in Education: Principles, Policy and Practices*, *6*(2), 177–194.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, *LI*, 599–635.

Eraut, M. (2004). A wider perspective on assessment. *Medical Education*, *38*, 803–804.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, *36*, 923–936.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.). *Educational measurement* (4th ed., pp. 65–110). Washington, DC: American Council on Education/Praeger.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27.

Hogan, T. P. (2007). *Educational assessment: A practical introduction*. Hoboken, NJ: John Wiley.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.). *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Kubiszyn, T., & Borich, G. (2015). *Educational testing and measurement* (11th ed.). Hoboken, NJ: John Wiley.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.). *Educational measurement* (4th ed., pp. 387–431). Washington, DC: American Council on Education/Praeger.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15–21.

Mari, L. (2000). Beyond the representational viewpoint: A new formalization of measurement. *Measurement*, *27*(2), 71–84.

Mari, L. (2013). A quest for the definition of measurement. *Measurement*, *46*, 2889–2895.

Maul, A. (2013). On the ontology of psychological attributes. *Theory and Psychology*, *23*, 752–769.

McCallin, R. C. (2006). Test administration. In S. M. Downing & T. M. Haladyna (Eds.). *Handbook of test development* (pp. 625–652). Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*, 215–237.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.

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

Messick, S. (1995a). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Messick, S. (1995b). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256.

Messick, S. (1998). Alternative models of assessment, uniform standards of validity. In M. Hakel (Ed.). *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 59–74). Mahwah, NJ: Lawrence Erlbaum.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed). Boston, MA: Pearson Education.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.). *Test theory for a new generation of tests* (pp. 19–39). Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36*, 463–469.

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

Newton, P. E. (2016). Macro- and micro-validation: Beyond the 'five sources' framework for classifying validation evidence and analysis. *Practical Assessment, Research and Evaluation*, *21*(12). Available online http://pareonline.net/getvn.asp?v=21&n=12

Newton, P. E. (2017a). *An approach to understanding validation arguments*. Ofqual/17/6293. Coventry, UK: Office of Qualifications and Examinations Regulation.

Newton, P. E. (2017b). There is more to educational measurement than measuring: The importance of embracing purpose pluralism. *Educational Measurement: Issues and Practice*, *36*(2), 5–15.

Newton, P. E. (2019). What is response process validation evidence and how important is it? An essay reviewing Ercikan and Pellegrino (2017) and Zumbo and Hubley (2017). *Assessment in Education: Principles, Policy and Practice*, *26*(2), 245–253.

Newton, P. E., & Baird, J. (2016). Editorial: The great validity debate. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 173–177.

Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and applications of mental appraisal* (2nd ed.). Boston, MA: Pearson Education.

Perie, M., & Huff, K. (2016). Determining content and cognitive demand for achievement tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 119–143). New York, NY: Routledge.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2010). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*(2), 191–209.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed). *Educational measurement* (4th ed., pp. 307–353). Washington, DC: American Council on Education/Praeger.

Schmidt, L., & Schuwirth, L. (2013). *Better judgement: Improving assessors' management of factors affecting their judgement*. Canberra: Australian Government Office for Learning and Teaching.

Schmidt, L., Schuwirth, L., O'Keefe, M., & King, S. (2016). *Better judgement—2: Improving assessors' management of factors affecting their judgement*. Canberra: Australian Government Office for Learning and Teaching.

Slaney, K. L., & Garcia, D. A. (2015). Constructing psychological objects: The rhetoric of constructs. *Journal of Theoretical and Philosophical Psychology*, *35*(4), 244–259.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing*, *2*(1), 31–40.

Thorndike, E. L. (1925). The improvement of mental measurements. *Journal of Educational Research*, *11*(1), 1–11.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.). *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson Education.

van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education*, *1*(1), 41–67.

van Fraassen, B. C. (2008). *Scientific representation*. Oxford, UK: Oxford University Press.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, *70*, 703–713.

Wiliam, D. (1992). Some technical issues in assessment: A user's guide. *British Journal of Curriculum and Assessment*, *2*(3), 11–20.

Wood, R., & Power, C. (1987). Aspects of the competence–performance distinction: Educational, psychological and measurement issues. *Journal of Curriculum Studies*, *19*, 409–424.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed,), *Educational measurement* (4th ed., pp. 111–153). Washington, DC: American Council on Education/Praeger.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*(2), 79–87.