Psicología Educativa 20 (2014) 89-97

Psicología Educativa

www.elsevier.es/psed

Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework

Jimmy de la Torre and Nathan Minchen

Rutgers, The State University of New Jersey, U.S.A.

ARTICLE INFORMATION

Manuscript received: 20/05/2014
Revision received: 22/09/2014
Accepted: 6/11/2014

Keywords:
Cognitively diagnostic assessments
Cognitive diagnosis model
Evidence-centered design
Assessment triangle

ABSTRACT

This paper aims to identify the utility of and the need for cognitively diagnostic assessments (CDAs) in conjunction with cognitive diagnosis models (CDMs), and to outline various considerations involved in their development and use. We begin by contrasting the CDA/CDM framework against existing assessment frameworks, which are typically based on item response theory or classical test theory, and show that CDAs used in the CDM context can provide valuable diagnostic information that could enhance classroom instruction and learning. We then detail how the components of a CDA fit into the assessment triangle framework, as well as the evidence-centered design framework. Attribute identification and item development in the context of CDA are discussed, and examples from relevant research are provided. Details of CDMs, which are the statistical models that underpin the practical implementations of CDAs, are also discussed.

El marco de la evaluación y los modelos de diagnostico cognitivo

Palabras clave:
Evaluación para el diagnóstico cognitivo
Modelo de diagnóstico cognitivo
Diseño centrado en la evidencia
Triángulo de evaluación

RESUMEN

El presente artículo trata de identificar la utilidad y la necesidad de las Evaluaciones para el Diagnóstico Cognitivo (EDC) junto a los Modelos de Diagnóstico Cognitivo (MDC) y plantea algunas de las consideraciones implicadas en su desarrollo y uso. Se comienza comparando el marco EDC/MDC con otros marcos existentes basados típicamente en la teoría de respuesta al ítem o en la teoría clásica de los tests, mostrando que las EDC utilizadas en el contexto MDC pueden proporcionar información diagnóstica muy valiosa para mejorar el proceso de enseñanza-aprendizaje en el aula. Seguidamente se utiliza el marco del triángulo de la evaluación y del diseño centrado en la evidencia para presentar los componentes de una evaluación de este tipo. Se analiza la identificación de atributos y el desarrollo de ítems en el contexto de una EDC y se proporcionan ejemplos tomados de una investigación relevante. También se analizan cuestiones relativas a los MDC, que son los modelos estadísticos que vertebran la implementación práctica de las EDC.

The Need for Cognitively Diagnostic Assessment

In traditional educational assessments, which are often rooted in item response theory (IRT) or classical test theory (CTT), a student's score is typically determined by identifying his or her location along a single proficiency continuum. With this particular interpretation, scores may be used as a part of a summative assessment program to compare or rank-order a student against other students, or against certain standards. Such scores can then be used for a variety of purposes, such as identifying a student's level of proficiency, differentiating passing from non-passing students, selecting candidates for a program, admitting students to a college, or determining the recipients of scholarships. Educational assessments used for these purposes are linked by their common goal of determining the extent to which students possess the proficiency or trait of interest.

These types of assessments fulfill an important function in the educational assessment and accountability landscape, and thus have been popularized in part due to their utility relative to these particular uses. However, it should be underscored that the original intent and design of these assessments do not naturally lend themselves to providing *diagnostic* information; thus these assessments do not provide sufficient diagnostic information that can be used to enhance classroom instruction and learning (de la Torre, 2009). Additionally, there can be a significant time lag between

*Correspondence concerning this article should be addressed to Jimmy de la Torre. Department of Educational Psychology Rutgers. The State University of New Jersey. 10 Seminary Place. New Brunswick, NJ 08901 USA. E-Mail: j.delatorre@rutgers.edu

DOI: http://dx.doi.org/10.1016/j.pse.2014.11.001
1135-755X/

the test event and the availability of results in large–scale testing. As a result, these types of assessments may not serve as catalysts for *immediate* change within the classroom; however, they may be of use in driving higher-level policy and programming decisions, such as the modification of curriculum and instruction for the *subsequent* years. Thus, there remains a need in educational measurement for assessments that can provide diagnostic information in a timely fashion; such assessments are the focus of this paper, and are referred to as cognitively diagnostic assessments (CDAs).

Because CDAs are fundamentally diagnostic, they require statistical models that are capable of extracting this level of information from the data. Such models are referred to as cognitive diagnosis models (CDMs) or diagnostic classification models (DCMs). Standing in contrast to the descriptive nature of IRT and CTT models, CDMs have been developed as an alternative psychometric framework to provide diagnostic information in the form of examinee classification with respect to a set of skills or attributes. CDMs can be viewed as constrained latent class models that model responses as a function of discrete latent variables (Templin & Henson, 2006). Thus, CDMs assume a collection of fine-grained attributes that are conceptualized as existing in a discrete latent space, and are typically, but not necessarily, binary, as opposed to the continuous latent ability continuum common in IRT and CTT models. However, it should be noted that CDMs are item response models applied to discrete latent variables. As such, with proper modifications, existing methods, procedures, and applications (e.g., estimation, model-fit evaluation, DIF analysis, computerized adaptive testing) in IRT can be easily adapted to apply in the context of CDMs.

For an assessment to be cognitively diagnostic, it needs to be designed to measure various components required of someone deemed proficient in a particular domain of interest. Such a design should allow for theories of learning, cognition, and pedagogy to be integrated with theories of measurement to develop assessments that not only measure, but also support student learning (Chudowsky & Pellegrino, 2003; Shepard, 2000). It is important to note that such a design presupposes a psychometric framework that is consistent with recent advances in cognitive theories and other relevant fields to address complexities inherent in academic learning (Pellegrino, Chudowsky, & Glaser, 2001). To this end, CDMs offer such a framework.

In a mathematical sense, CDMs can measure an unlimited number of attributes, although a good rule of thumb for an upper limit is 10 attributes, due to the number of possible combinations of items possible (Tatsuoka et al., in press; DiBello, Roussos, & Stout, 2007), especially if their structure cannot be ordered hierarchically (i.e., some attribute combinations cannot be discounted a priori). This can make administration of assessments that measure a large number of attributes impractical because the maximum number of possible attribute combinations can grow exponentially with the number of attributes. Conversely, assessments based on IRT are often uni- or low-dimensional (Junker & Sijtsma, 2001), and consequently are of limited value to inform classroom instruction and learning because they only assess a small number of, if not highly correlated, general abilities. Due to the finer grain-size that typically characterizes attributes used in CDAs, a continuous scale no longer makes sense; rather, attributes are characterized as being either present in or absent from examinees. In the educational measurement context, these labels are generally thought of as mastery and non-mastery of a particular domain. (In the medical or psychological setting, these labels can be thought of as a patient either meeting or not meeting a criterion for diagnosis.) Test items in CDAs can be designed to measure a single attribute or combinations of attributes in examinees.

In contrast, the coarser-grained scores from tests that measure a continuous proficiency provide theoretical location or ability estimates of students along a scale continuum, which can be useful, but may or may not directly translate into actionable steps teachers can take to adapt and respond to students' needs. Despite their limitations, IRT-based tests have been used for diagnostic purposes (de la Torre & Karelitz, 2009; de la Torre, 2012). This is in stark contrast to the very nature of the theory underpinning CDMs, which lends itself to diagnostic purposes. For example, scale scores from an IRT-based test can be used to locate an examinee on the ability continuum, which is the same continuum that defines the difficulty of the items. Based on the estimated location, broad inferences can be made about the types of problems a student can answer correctly, which is done through item-mapping. However, as de la Torre (2012) noted, the difficulty of an item "is a coarse summary of the different features that make an item easy or difficult" (p. 4). In other words, unless these various features are disentangled, it will be difficult to provide specific direction on which actual aspects of the domain the student has mastered.

An example of the item mapping process is given in Figure 1, which shows the partial item map for three selected scores for the 2013 NAEP Grade 8 Mathematics test (U.S. Department of Education, 2013). Each of these scores corresponds to a different proficiency level: 331 is at the top of the Proficient level (just below Advanced), 296 is at the top of the Basic level (just below Proficient), and 257 is below Basic. These descriptions help to exemplify the varying skill levels associated with their respective scale scores. Although these descriptions provide some information about what types of problems students can and cannot solve, they are quite specific and lack an overall theme or cohesive theory, which is consistent with de la Torre's (2012) remark. In other words, this information may allow a teacher to focus on one particular task, but that task may grossly underrepresent the actual domain deficiency.

In this way, scale scores may be capturing elements of items that are unique to the items themselves and not representative of a well-defined domain or attribute. This means that score reports that provide information on the types of problems a student can solve may be unreliable or misleading (de la Torre, 2012). Also, consistent with the limited dimensionality and continuous ability scale of IRT and CTT models, the proficiency or proficiencies measured, such as Grade 8 Mathematics in this example, are generally broader resulting in a grain-size that is coarser. Without detailed guidance on how these relatively coarse test scores should be interpreted and used to inform instruction and learning, reliance on individual interpretation can inject subjectivity that can result in serious misuse and/or misinterpretation of test scores.

Moreover, without a deliberate attempt to keep the interpretation of different test scores across the range consistent, the interpretation of test scores is liable to change when similar or modified problems are administered over multiple time points. For example, when different numbers or options are associated with one's ability to calculate the areas on an inscribed square in a future test administration, this description may no longer be appropriate to characterize proficient students.

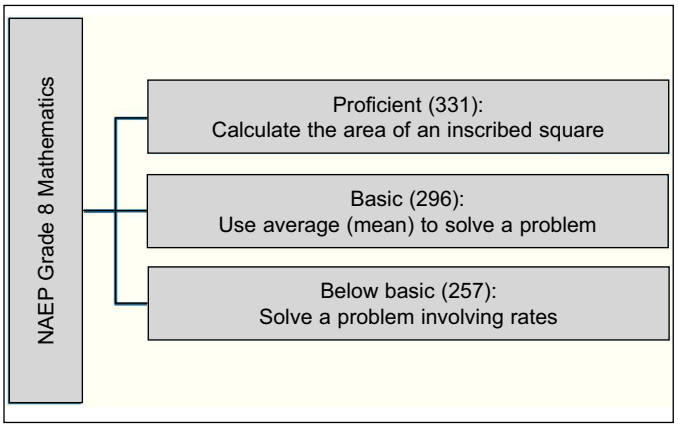

Figure 1. Partial Item Map for 2013 NAEP Grade 8 Mathematics Test

Because CDMs are a type of latent class model, they are designed to classify students rather than to order them by rank. There are certain exceptional situations, though, where latent classes could result in ranking (i.e., when only subset of classes is permissible and they can themselves be ranked), and this will be briefly discussed in the context of retrofitting later in this article. However, although there exists a vast array of assessments designed to rank order students on the proficiency of interest, it is important to note that CDMs are generally not intended to be used with these assessments.

When the construction of CDAs is guided by the relevant cognitive and substantive considerations – which will be discussed next – one of several CDMs can be employed to obtain a classification of the examinees. The end result of such an assessment is that examinees are classified into groups according to the combination of attributes they have mastered. At the group level, the proportion of examinees that possess a particular attribute can also be estimated. In the context of education, this classification, which is based on the student profiles, allows teachers to have a richer understanding of what specific combinations of skills have been mastered among their students. Teachers can use the information from the reported student attribute profiles to tailor subsequent instruction, providing the double benefit of serving students and maximizing the efficiency of classroom instructional time. Increased efficiency and targeted instruction could be expected to lead to greater content coverage and increased understanding, which in turn could result in better performance on summative exams. This, of course, remains purely a conjecture at this point. Proper construction of a CDA is by no means a trivial matter, and will be discussed in detail in the next section.

An Evidence Centered Design Approach to Cognitively Diagnostic Assessment

In this section, we follow two well established organizing frameworks to demonstrate the distinct role and unique contributions of a CDA in the educational assessment process. The first of these frameworks is the *assessment triangle* discussed in *Knowing What Students Know,* which represents the process of *reasoning from evidence* (Pellegrino et al., 2001). The second is the evidence-centered design (ECD) framework, which was outlined by Mislevy, Steinberg, and Almond (2003). DiBello et al. (2007) argue that ECD provides a framework for the assessment triangle that results in an assessment design that can be implemented in practice, underscoring the commonalities of the frameworks. Furthermore, both of these can be viewed as process models that allow observed performances to be explained by the underlying processes (Kane, 2013).

The assessment triangle is comprised of three corners that represent the components in the process: the *cognition* component, the *observation* component, and the *interpretation* component. Briefly, the cognition component describes the theory of learning that specifies how knowledge and competency is obtained and displayed; the observation component specifies the types of tasks or activities which would display the knowledge or competency specified in the cognition component; and finally, the interpretation component represents our imperfect attempt to translate student responses (i.e., observations) into useful information (Pellegrino et al., 2001).

These corners share a dynamic relationship in which one corner informs the others, so they must be considered together. For example, the relevant cognitive or substantive theory can inform the nature and types of attributes to be measured, and the types of tasks needed to demonstrate mastery of these attributes. Similarly, the method of interpretation (i.e., CDM) can inform how tasks can be designed for the resulting observations to be amenable to proper interpretation. A process that involves multiple and iterative refinements and redefinitions may be needed before consistency between the three elements of the assessment triangle can be achieved (Pellegrino et al., 2001). In the same manner, ECD also may follow an iterative process (Mislevy et al., 2003).

Thus, the inferences that the assessment designer intends to draw from the collection of observations is guided by an interactive and iterative reasoning process through which a judgment is made relative to the purpose of the assessment. Any statistical models used to make this linkage are contained in the interpretation component. For example, in an intelligence test, a statistical model would take the responses as input and would provide an estimate of intelligence as the output. The assessment triangle is of use because it provides a general and overarching paradigm within which the components of CDA development are organized.

ECD is a framework that "entails the development, construction, and arrangement of specialized information elements, or assessment design objects, into specifications that embody the substantive argument that underlies an assessment" (Mislevy et al., 2003, p. 4). The authors outlined an overview of the structures of ECD as consisting of domain analysis, domain modeling, conceptual assessment framework, and operational assessment, the first three of which will be discussed in this paper. As mentioned earlier, DiBello et al. (2007) discussed ECD as framing the assessment triangle. We would add to this observation that ECD provides necessary specificity to each corner of the triangle. Thus, our approach will be to situate the relevant pieces of ECD beneath the appropriate corner of the assessment triangle, which does not assume a hierarchy; rather, we agree with remarks of DiBello et al. (2007) in that ECD presents a pragmatic place to begin the assessment design process because it outlines actionable steps.

Therefore, the purpose of this section is to outline the major components in developing a valid CDA using both the assessment triangle and ECD. (For a more general overview of assessment triangle and ECD and how they relate to other frameworks, see Pellegrino and Zieky, this volume.) Despite our systematic presentation, it should again be emphasized that the process of assessment design is not necessarily linear (DiBello et al., 2007). For example, although attribute definition typically precedes task construction, the feasibility of constructing tasks that unambiguously measure a particular set of attributes may require revisiting how the attributes have been defined. In this scenario, task construction informs attribution definition, which in turn will inform the next cycle of task construction. In this paper, whenever applicable, examples from empirical research studies will be presented.

Cognition

Domain analysis. The goal of the domain analysis layer of ECD is to compile and analyze information about the domain of interest, attending to the nature of the construction of knowledge (Mislevy 2011; Mislevy & Haertel, 2006; Mislevy et al., 2003). Therefore, the CDA building process begins by carefully and thoroughly examining the domain of interest. In working on a proportional reasoning assessment for middle school students, Tjoe and de la Torre (2013a, 2013b) provide a meticulous documentation of this process, including direct observation of student reasoning. Thus, the sections that follow will use their works as an example to illustrate the process in practice.

It may be tempting for researchers or educators to apply a CDM to data that arises from unidimensional IRT-based tests in an effort to capitalize on the benefits that the CDA/CDM framework offers. This practice is known as *retrofitting*. de la Torre and Karelitz (2009) found retrofitting to be suboptimal, often resulting in the misclassification of examinees. Therefore, we generally discourage this practice due to weak or nonexistent theoretical justification, except perhaps for exploratory and research purposes.

Purpose. From the purpose of the assessment stems the identified domain(s) which will comprise the assessment, which are referred to as *targets of inference* (Pellegrino et al., 2001). Therefore, defining the purpose of the assessment is vitally important (DiBello et al., 2007), and is a natural first step in the domain analysis. The purpose of an assessment may be developed in a local education community, or it may be imposed from higher authorities, such as state/provincial and federal governing bodies, or it may be developed by a testing agency or researchers in response to legislation or perceived need, but it is critical because it will guide the remaining steps in the process. The stated purpose serves to direct not only the development of the actual instrument, but also the interpretation/use argument (IUA), which is the argument built to justify the intended interpretation and use of the test scores (Kane, 2013). Establishing and providing support for the IUA is a critical component of the overall validation of an assessment system. Thus, a clearly articulated purpose needs to be specified and will help researchers and experts begin to preliminarily identify the construct(s) of interest and the number and type of resulting attributes that will be measured. The final product, which is the instrument itself and the diagnostic outcomes reported to educators, is, in part, the result of this step.

Identifying attributes. Once the targets of inference have been identified, the next step in creating the assessment is to identify what will be measured: the attributes (DiBello et al., 2007). The attributes arise out of the cognitive model specified in the cognition corner of the assessment triangle, and should "reflect the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise in a domain" (Pellegrino et al., 2001, p. 45). These findings should be based on sound research in education and cognition, as well as the expertise of experienced and highly qualified teachers (Pellegrino et al., 2001). The core of the domain analysis process lies in the gathering of evidence necessary to identify a collection of suitable attributes.

Due to expedience and other practical constraints, some researchers may identify attributes using other approaches. One such approach may be to build the attribute space for an existing assessment, representing an ex post facto approach to determining attributes. However, solely relying on existing items is essentially a retrofitting strategy that can limit the scope of what attributes can be measured, which is partly due to the fact that this is essentially a post-hoc domain analysis with no clearly defined purpose. Such a strategy, which is not based on a strong understanding of cognition and learning as is outlined here, may render assessments theoretically impoverished and practically useless. Furthermore, there is no way to ensure that the measurement of the attributes is balanced.

For items to be created, it is first necessary to identify and define the characteristics of attributes; their characteristics, such as their number and grain-size, will to a large degree determine the breadth of content coverage that is appropriate for a given assessment. Tjoe and de la Torre (2013b) enumerated several important characteristics that are desirable in attributes. First, attributes need to be common to many problems in the subject area. They also need to be of an appropriate grain-size, which means that they will be appropriate for the diagnostic purposes of the assessment. As an illustration, the *proportional reasoning* ability defined by Tjoe and de la Torre (2013b), which subsumes multiple subabilities, has a sufficiently large grain-size that it can be viewed as a continuous ability. In contrast, the attribute *ordering fractions*, which represents a component of proportional reasoning, has a small enough grain-size for it to be construed in a dichotomous manner (i.e., student can master or not master the attribute). In a sense, an attribute is a rudimentary component that sits alongside other components of a similar primacy to a larger domain. Tatsuoka et al. (in press) indicate that the level of detail increases as the grain-size decreases, necessarily limiting the scope of the content coverage. In addition to the inverse relationship that suggests that the selection of the grain-size is dependent upon the depth and breadth of information to be extracted from the assessment, other factors, such as the practical testing constraints and alignment with instructional and assessment practices can affect the granularity by which attributes are defined.

In summary, the domain analysis component of the ECD framework captures many of the elements of the cognition corner of the assessment triangle. In the domain analysis component, we investigate the appropriate nature and characteristics of the attributes with respect to the purpose of the assessment itself. This is analogous to identifying the construct(s) that are typically referred to as Θ in IRT-based assessments. At this stage, this is largely a theoretical pursuit; testing or validating the findings of this step is the primary goal in the domain modeling step, which is the next step to be presented following the example.

Example. Tjoe and de la Torre (2013b) began the process of identifying attributes with an extensive literature and database review. This allowed the authors to be informed about the current research in proportional reasoning at the middle school grade level. This preliminary literature review was important because it prepared the authors for the next phase in the project, in which experts in the field of mathematics –researchers, educators, and middle school teachers– were consulted (Tjoe & de la Torre, 2013b). A series of meetings ensued. The goal of the first meeting was to begin the process by determining which attributes were both of some value to researchers and practitioners, and also amenable to assessment (in a psychometric sense).

Table 1
Proportional Reasoning Attributes

Attribute	Description
A1	Prerequisite skills and concepts required in proportional reasoning
A2a	Comparing (two) fractions
A2b	Ordering (three or more) fractions
A3a	Constructing ratios
A3b	Constructing proportions
A4	Identifying a multiplicative relationship between sets of values
A5	Differentiating a proportional from a non-proportional relationship
A6	Applying algorithms in solving proportional reasoning problems

Participants were asked to enumerate and define a list of skills (Tjoe & de la Torre, 2013b) that would serve as a preliminary list of attributes. In the next meeting, this list was whittled down to six attributes, and in a final meeting, meaning and measurability issues were tackled. The list of proportional reasoning attributes they arrived at is given in Table 1. In should be noted that Attributes 2 and 3 have a hierarchical structure, in that the "a" component must be mastered to master the "b" component.

Observation

Domain modeling. In the domain modeling layer of ECD, "designers organize information from domain analyses to describe relationships among [the] target knowledge and skill, what we might see people say, do, or make as evidence" (Mislevy, 2011, p. 9). In domain modeling, detailed statements are needed to articulate the types of evidence required to reasonably infer student knowledge in the domain. A clear understanding of the relationship between evidence and tasks can provide the necessary guidance on how the tasks must be designed to elicit the evidence of interest.

In the context of CDA, initial tasks based on the desired evidence can be constructed. By design, each task is constructed to measure different combinations of attributes. However, one cannot simply

take for granted that the tasks individually and collectively will actually measure the attributes as intended. For this reason, the domain-modeling layer involves an additional process of gathering validity evidence that supports the use of the attributes, which eventually contributes to the evaluation of the IUA for the assessment.

In general, the process of validation is to seek empirical evidence that supports a theory; validation in this context is no different. In the context of CDA, Tatsuoka et al. (in press) identified that the critical task of the validation process is to verify that examinees are employing the hypothesized attributes, and those attributes only, to correctly solve each problem. To establish this kind of validity evidence requires an observation process that elicits the skills one is using or failing to use in problem solving.

Validation example. Tjoe and de la Torre (2013b) discussed a number of activities that were pursued to ensure that the attributes of interest were indeed being measured. One critical component of the validation process is documenting how different researchers solve proportional reasoning problems. Another critical component of the process was to examine how students solve the same problems. Using a think-aloud protocol, insights on the students' thinking and problem-solving processes were collected and analyzed. Based on this work, Tjoe and de la Torre concluded the viability of the six attributes given in Table 1. The same data also served as a way to inform the content of distractor items, which sometimes contained the most common mistakes (Tjoe & de la Torre, 2013a).

Interpretation

Conceptual Assessment Framework (CAF). The CAF layer is essentially the creation of a blueprint for an assessment, which is developed using the domain information and other information related to the goals of the assessment and constraints under which it needs to operate (Mislevy, 2011). As such, we will examine three components of the CAF: the *student model*, the *evidence model*, and the *task model*.

In the student model, variables are defined that appropriately account for and theoretically conceptualize the observations. In this example, the end result of the domain analysis and domain modeling processes –the validated list of attributes– serves as the student model; it is both what is measured and what is reported. Score reports for CDAs can specify the probability that an examinee possesses a given attribute. In addition, extended definitions of the attributes and exemplar problems can help teachers better understand the nature of the attributes being measured, and how classroom instruction can be designed or modified to better facilitate the mastery of these attributes.

In the task model, a formal description of the test environment, which specifies the ways that evidence will be collected from examinees, is produced (Mislevy, 2011). The task model consists of many of the details of the assessment. A significant component of the task model is the format of the test. Examples of formats include multiple-choice, open-ended, and teacher-constructed tests. An example of a task model is the proportional reasoning assessment item given in Figure 2. In this item, a single attribute is being measured: constructing ratios (A3a).

A farm has 8 cows, 15 sheep, and 6 goats. What is the ratio of cows to sheep?

A. $\frac{5}{3}$
B. $\frac{8}{6}$
C. $\frac{8}{15}$
D. $\frac{6}{18}$

Figure 2. Sample Proportional Reasoning Item

Tjoe and de la Torre (2014) argued that developing CDAs may involve writing novel item types. Their previous research showed that missing value problems (MVPs) are the most common type of proportional reasoning problems (Tjoe & de la Torre, 2012). However, as they are, MVPs require multiple attributes (i.e., A1, A3, A5 and A6), making it difficult for students' strengths and weaknesses to be isolated. In addition, Tjoe and de la Torre (2012) found that the ability to solve MVPs does not necessarily imply the ability to differentiate a proportional relationship from one that is non-proportional.

The example given in Figure 3 demonstrates a new type of problem that allows a proportional reasoning attribute (i.e., A5) to be isolated. In this example, a student cannot carelessly apply an algorithm to solve for any missing value as will suffice in most MVPs; rather, he or she must identify the situation(s) where the given proportion can be used appropriately, which represents a unique skill that does not result from combining the other attributes. In other words, if a student was in possession of all other skills in this example with the exception of A5, there is no theoretical basis to believe that a student should correctly respond to this problem.

The proportion 1/2 = 10/x can be used to solve which of the following situation?

Situation I: For every 1 boy, there are 2 girls in a classroom. If there are 10 boys in the classroom, how many girls are there?

Situation II: Bob is 1 year old and Mary is 2 years old. When Bob is 10 years old, how old will Mary be?

Figure 3. A Proportional Reasoning Problem Measuring A5

The evidence model provides the link between the student and the task models, and it consists of two components: the *evaluation component* and the *measurement component*. This is perhaps the piece of ECD that most closely relates to the interpretation corner of the assessment triangle. The *evaluation component,* which is the method of scoring tasks, essentially details how observations will be evaluated and expressed in statistical form (Mislevy et al., 2003). In the case of this example, the format of the test is multiple-choice, and there is only one correct answer for each question. Therefore, examinee responses are recorded as correct or incorrect and are converted to binary vectors of length *J*, which represents the number of items on the test. An alternative to a binary response is a polytomous response, in which categories can either be conceptualized as nominal (e.g., type of misconceptions) or ordinal (e.g., partial credit). More generally, the type and purpose of the item will dictate the evaluation component necessary. For example, items with a constructed response may require rubrics for scoring.

The *measurement component* is the statistical model employed to analyze response patterns whose function is to link the latent student ability to the observed variable. Specifically, the measurement component is of value because it is the mechanism that allows for *reverse reasoning* (Mislevy et al., 2003) to systematically estimate the status or level of a latent variable (ability) from the value of an observed variable (performance). Well known examples of measurement models are IRT, CTT, and, more recently, CDMs. Measurement models allow for large amounts of testing data to be handled efficiently, practically, and reliably. Particularly for formative and diagnostic purposes, where time can be of the essence, this is of critical importance. We now turn to a short didactic in which several specific CDMs are described in detail.

Cognitive Diagnosis Models

As stated earlier, CDMs are generally employed with the purpose of measuring a collection of finer-grained attributes or skills; thus it is necessary to introduce some basic notation to represent these

models in mathematical form and to discuss their characteristics. First, attributes are characterized as being discrete and dichotomous, and thus, they are either present in or absent from an examinee or an item. The combination of skills that an examinee possesses is defined as an attribute pattern, $\boldsymbol{\alpha}$, and is represented as a latent K-length vector in which a 0 or 1 in the kth entry represents nonmastery and mastery of the kth attribute, respectively. Similarly, each item has a corresponding q-vector, $\boldsymbol{q}$, which is also of length K, and represents the attributes that are required to solve that item. If the kth attribute is required to solve the jth item, $q_{jk} = 1$, otherwise $q_{jk} = 0$. Collecting all the q-vectors in a test of length J results in a Q-Matrix (Tatsuoka, 1983) of dimension $J \times K$. It should be noted that, whereas $\boldsymbol{q}$ is directly observable after construction, albeit potentially with misspecifications, $\boldsymbol{\alpha}$ is not and needs to be estimated. For instructional and learning purposes, the chief goal of cognitive diagnosis modeling is to accurately estimate $\boldsymbol{\alpha}$ or, equivalently, classify students into one of the potentially 2^K attribute combinations, each of which represents a unique latent class (i.e., unique mastery and nonmastery pattern). It is important to emphasize that the inferences about students are only valid to the extent that the list of attributes is complete, the attributes are correctly associated with the items (as specified in the Q-matrix), and the items are relevant to the attributes being measured.

The DINA Model

A variety of specific CDMs have been developed, each of which assumes a particular cognitive model that specifies the nature of the interaction between examinees' attributes and the attributes in the items. We begin by examining the *deterministic inputs, noisy "and" gate* (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model, which is one of the simplest CDMs. CDMs assume a relationship between the examinees' attributes and the skills required to solve a problem. In the case of the DINA model, it is assumed that examinees must possess all the skills required to effectively solve an item, making it a conjunctive model; lacking one required attribute cannot be made up for by the presence of other attributes.

Consequently, the DINA model partitions examinees into two latent groups for each item. In group 1, examinees have all required attributes to solve item j, and in group 0, examinees lack at least one of the required attributes to solve item j. Thus, examinees lacking one attribute are considered the "same" as examinees lacking all attributes. Figure 4 shows an example of the probabilities of success on an item based on an item that requires three attributes. The figure shows that examines who lack one, two, or three required attributes have an identical success probability of .2: only examinees who possess all the required attributes have the maximum probability of .8.

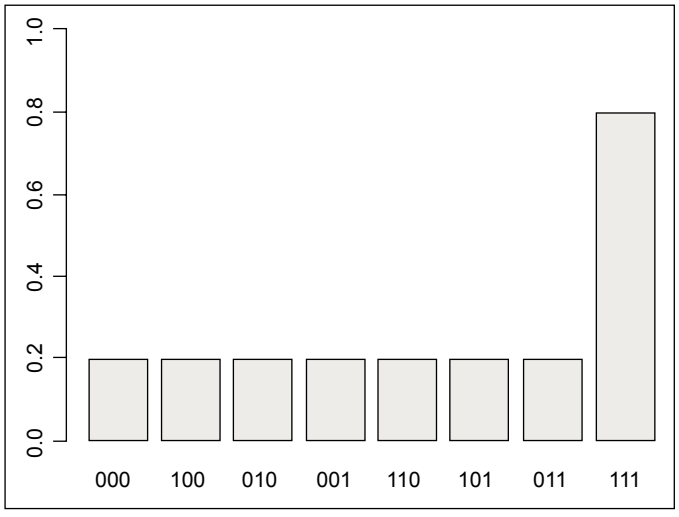

Figure 4. The DINA Model

The interaction between the examinee attributes and the item specification defines the latent response variable, which is also known as the *ideal response* (Tatsuoka, 1995). The ideal response is defined as

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ij}^{q_{jk}} \tag{1}$$

for the DINA model. Thus, $\eta = 1$ if and only if examinee i has mastered all the required attributes for item j. To account for the probabilistic nature of the observed response, slip and guessing parameters conditional on the ideal response are defined at the item level. These parameters are simply "false negative and false positive rates" (Junker & Sijtsma, 2001, p. 263). In this context, false negative refers to the probability that an examinee with all required attributes provides an incorrect response. Conversely, false positive refers to the probability that an examinee without all required attributes provides a correct response. The slip and guessing parameters are given by

$$s_j = P(X_{ij} = 0 \,|\, \eta_{ij} = 1), \text{ and} \tag{2}$$

$$g_j = P(X_{ij} = 1 \,|\, \eta_{ij} = 0), \tag{3}$$

respectively. Note that $1 - s_j = 1 - P(X_{ij} = 0/\eta_{ij} = 1) = P(X_{ij} = 1/\eta_{ij} = 1)$, and thus, the response function for an item is given by

$$P(X_{ij} = 1 \,|\, \eta_{ij}) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} \tag{4}$$

The DINA model can be seen as being too simple and restrictive (de la Torre, 2011; Henson & Douglas, 2005), and thus may only have limited applicability. However, conjunctive models may be specifically required in settings where it is important that examinees have all required skills. An example of a conjunctive process that may follow the spirit of the DINA model is an open-ended problem that does not provide partial credits. If the problem requires three attributes, examinees who master two of the three required attributes will not score higher than someone who has not mastered any of these attributes. Note that their probability of success may change once the test format is changed. In a multiple-choice test, examinees who can perform two of the three required steps may have a higher probability of guessing the correct response than someone who is guessing completely at random.

De la Torre and Lee (2010) noted that the DINA model parameters are absolutely invariant. This property, which also applies to other CDMs, allows calibration of item parameters without requiring arbitrary constraints to be set. For example, constraints such as setting the mean and standard deviation of the proficiency distribution to 0 and 1, respectively, are typically used in IRT. Thus, so long as the model fits the data, CDM item parameter estimates, and hence, examinee classifications, are comparable without the need to explicitly equate them.

The G-DINA Model

We now introduce the *generalized* DINA (G-DINA; de la Torre, 2011) model, which is a generalization of the DINA model to address the strong conjunctive assumption asserted in the DINA model. In the DINA model, all examinees lacking one or more of the required attributes for an item each have the same probability of success regardless of how many attributes they have, which for the DINA model is simply the guessing parameter. With the G-DINA model, this constraint is removed, and each group of examinees can have its own probability of success. Figure 5 shows a possible arrangement of the probabilities of success on a three-attribute item for each latent group.

Whereas the DINA model partitions examinees into two groups regardless of the number of required attributes, the G-DINA model partitions examinees into $2^{K_j^*}$ groups, where K_j^* is the number of attributes required for the jth item. For example, when K_j^* =3, instead of just 2, there will be 2^3 = 8 latent groups created for the item using the G-DINA model, all of which are free to have different probabilities of success. It should be noted that when K_j^* =1, the DINA and G-DINA are one and the same model.

Assume that the first K_j^* attributes are required for item j. In addition, define $\boldsymbol{\alpha}_{ij}^*$ as the attribute vector comprising the first K_j^* elements of $\boldsymbol{\alpha}_i$. The item response function for the G-DINA model can be defined as

$$P(\boldsymbol{\alpha}_{ij}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ik} + \sum_{k=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{ik'}\alpha_{ik} + \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}, \quad (5)$$

where δ_{j0} is the intercept, δ_{jk} is the main effect due to α_k, δ_{jkk^*} is the two-way interaction effect due to α_k and α_k and $\delta_{j12\ldots K_j^*}$ is the K_j^*-way interaction effect due to α_1 through $\alpha_{K_j^*}$. The intercept can be thought of as the baseline success probability when no required attributes are present, the main effect as the change in success probability when one attribute is mastered, and the interaction effects as the change in success probability when more than one attribute is simultaneously mastered.

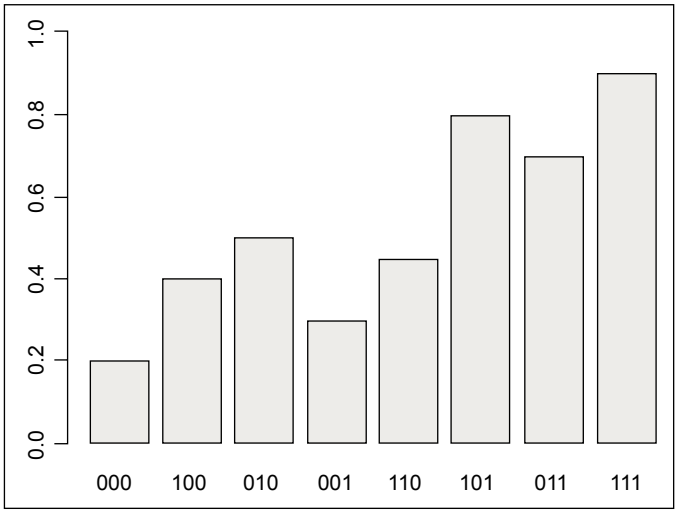

Figure 5. The G-DINA Model

As a saturated model, the G-DINA model includes all possible interaction terms. These interaction terms make it possible for an examinee with more required skills than another examinee to actually have a lower probability of success on an item. This interactive effect can be seen in Figure 5. Group 010 actually has a higher probability of success than group 110, indicating that the addition of an attribute does not always result in a higher probability of success.

Other CDMs

Various constraints may be applied to this model to suit the needs of the researcher. If all interaction effects are constrained to be 0, the additive CDM (*A*-CDM; de la Torre, 2011) is obtained:

$$P(\boldsymbol{\alpha}_{ij}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ik} \cdot \quad (6)$$

The *A*-CDM assumes that mastering a required attribute has a contribution to the probability of success that is independent of the contributions of other attributes. It should be noted that other additive models (i.e., linear logistic model, Maris, 1999; reduced reparameterized unified model, Hartz, 2000) exist, albeit based on different link functions.

Another CDM that can be derived from the G-DINA model is the *deterministic input, noisy "or" gate* (DINO; Templin & Henson, 2006) model. Similar to the DINA model, the DINO model partitions examinees into two groups for each item: those who have none of the required attributes, and those who have at least one of the required attributes for success. In this way, the DINO model is a disjunctive and completely compensatory model because an examinee can compensate for lacking a skill by possessing another skill. The ideal response variable for the DINO model is defined as

$$\nu_{ij} = 1 - \prod_{k=1}^{K} (1-\alpha_{ik})^{q_{jk}} \quad (7)$$

The item response function for the DINO model conditional on the latent response variable ν_{ij} is defined as

$$P(X_{ij} = 1 \mid \nu_{ij}) = (1-s_j)^{\nu_{ij}} g_j^{1-\nu_{ij}} , \quad (8)$$

where s_j and g_j are the slip and guessing parameters, respectively. Although Templin and Henson (2006) developed the DINO model to address diagnosis of a psychological condition, this model can also be applied in educational assessment. Specifically, if an attribute represents a particular strategy, then the DINO model can be used to model multiple-strategy problems.

De la Torre (2011) showed that the *A*-CDM can be readily derived from the G-DINA model. Similarly, the DINA and DINO models can be obtained as special cases of the G-DINA model.

It should also be noted that multiple-choice CDMs exist to go beyond right-or-wrong (i.e., 0/1) scoring (see de la Torre, 2009). However, this is a different type of measurement model, and is beyond the scope of the example at hand. Because these models require more information from the data, employing a multiple-choice model would require adjusting the *evaluation component* of the evidence model to produce a vector that is polytomous, albeit unordered, rather than dichotomous in nature.

Conclusion

IRT- and CTT-based summative assessments would not be virtually ubiquitous if it were not for their merits. They fulfill many vital functions in the assessment landscape. However, no single type of test can serve disparate demands and expectations on assessments. For various reasons, summative assessments cannot be effectively used for formative purposes. For assessments to be relevant in practical classroom settings, they need to be deliberately designed for this purpose. As argued in this paper, the CDM framework can be used to design and develop CDAs that are diagnostic and can inform classroom instruction and learning. However, critical as CDAs may be, they are just the first step in realizing the diagnostic potential of assessments. To complete the process, appropriate tools (i.e., CDMs) need to be employed to harness the diagnostic information available in these assessments. Both CDAs and CDMs are needed as they complement each other.

Due to the prevalence of applications that retrofit CDMs to extant test data, it is worthwhile to reiterate that such a practice should be discouraged whenever possible. However, we also acknowledge that in some cases retrofitting CDMs may be acceptable. Leighton, Gierl, and Hunka (2004) noted that there were only a few specific exceptions when misclassification of examinees does not occur, one of which is when the attributes have a linear hierarchical attribute structure because the presence of each additional attribute can be mapped onto a specific point on the proficiency continuum. A caveat could be added, however, that under this condition, a unidimensional model may provide equal, if not, a better fit.

At present, developments in CDMs far outpace developments in CDAs. Although other issues in CDMs remain to be solved (e.g., validating attributes and Q-matrices more efficiently), the essential methodological infrastructures are already in place. Therefore, the time is ripe for attention and resources to be focused on designing and developing educational assessments that are based on a CDM framework. To accomplish such a feat would require enjoining experts from different fields (i.e., subject matter, learning sciences, measurement, pedagogy) to work together. In doing so, we can add value to educational assessments and make them more relevant to the needs of present day classrooms.

Resumen ampliado[1]

Las evaluaciones al uso suelen tener el cometido de determinar cuál es el nivel de competencia en un determinado campo de un estudiante (o de un grupo de estudiantes). Para ello se suele recurrir a la Teoría Clásica de los Tests (TCT) o a la Teoría de Respuesta al Ítem (TRI), que ofrecen habitualmente una estimación de dicha competencia representada mediante un punto en una escala continua; el valor obtenido se utiliza para ordenar o comparar a los estudiantes (o grupos) entre sí o bien con algún criterio o estándar previamente definido con objetivos diversos (cualificación de la competencia, admisión a un programa, obtención de becas, acceso a la Universidad, rendición de cuentas).

El carácter continuo de la escala utilizada por estas dos teorías de tests junto a la dimensionalidad –habitualmente baja– de las pruebas con las que se suele operar en ambos modelos lleva a trabajar con estimaciones de la competencia que pueden servir muy bien a los objetivos anteriores, pero que resultan más bien toscas para poder contribuir a mejorar el proceso de enseñanza-aprendizaje en el aula: difícilmente se pueden traducir en medidas concretas o pasos a seguir por el profesor para adaptarse y responder en clase a las necesidades de sus estudiantes, ya que no proporcionan información lo suficientemente diagnóstica, por la propia naturaleza y objetivo de estas evaluaciones sumativas basadas en la TCT o en la TRI.

En la Evaluación para el Diagnóstico Cognitivo (EDC) se definen y miden atributos o variables que son relevantes para la práctica en el aula, promoviendo de este modo el aprendizaje de los alumnos. Para ello se trabaja integrando las teorías cognitivas con el corpus teórico procedente de otros campos de interés para abordar la complejidad intrínseca del aprendizaje académico, utilizando una teoría de tests acorde, esto es, los modelos de diagnóstico cognitivo.

Los autores utilizan el marco del triángulo de la evaluación y del Diseño Centrado en la Evidencia (DCE) para presentar los principales componentes de una evaluación de este tipo.

En el vértice de la cognición del triángulo de evaluación tienen lugar las actividades propias de la primera etapa (o capa) del DCE: el análisis del dominio de interés sobre el que se va a trabajar en la evaluación en cuestión, que conducirá a expresar de forma muy clara el objetivo del test que se va a construir (las inferencias básicas que se desea realizar en base a las puntuaciones del test), lo que se va a medir (los atributos de interés procedentes de la correspondiente teoría cognitiva que sustenta el vértice de la cognición) y la utilidad formativa de los resultados del test para los profesores. En suma, en esta etapa se construye el denominado argumento de interpretación/uso de Kane (2013). Los autores insisten en la necesidad de diseñar específicamente el test para el propósito deseado y evitar la práctica de ajustar a posteriori un modelo de diagnóstico cognitivo (*retrofitting*) a los datos obtenidos al administrar un test unidimensional construido según los dictados de modelos de más baja dimensionalidad como la TCT o la TRI.

En el vértice de la observación del triángulo de evaluación se comienza a construir el argumento de validez en la etapa del DCE denominada modelado del dominio: hay que determinar qué tipo de evidencia se necesita para poder realizar con garantías las inferencias deseadas acerca del conocimiento de los estudiantes del dominio en cuestión y ver también si éstos están utilizando o no los atributos o características hipotetizadas. La información obtenida en esta etapa ha de contribuir a diseñar tareas que sirvan para obtener esa evidencia.

En el vértice de la interpretación del triángulo de evaluación se construye el modelo de estudiante, el modelo de tarea y el modelo de evidencia en la tercera etapa del DCE, conocida como marco conceptual de la evaluación. Los autores se centran en el modelo de evidencia y, en particular, en su componente de medición, esto es, en el modelo estadístico que analiza las respuestas que los estudiantes han dado a las preguntas del test y que pone éstas en relación con el conjunto de atributos medidos. Los modelos de elección en el marco de una evaluación para el diagnóstico cognitivo son en buena lógica los Modelos de Diagnóstico Cognitivo (MDC).

Los MDC son modelos de clase latente –no de rasgo latente como la TRI (o la TCT)– que evalúan más atributos que las teorías anteriores y atributos con un menor nivel de generalidad que éstas, definiendo un espacio latente discreto –habitualmente binario– en el que se estima la probabilidad de que un estudiante presente o no cada uno de esos atributos. De este modo, se obtiene información diagnóstica en forma de clasificaciones de los estudiantes en función de la combinación de atributos que éstos muestran. Estas clasificaciones basadas en el perfil de los estudiantes proporcionan a los profesores información útil y práctica que les puede servir para adaptar la instrucción a cada estudiante individual o a la composición de la clase.

Los autores presentan con algún detalle el modelo más simple de todos (el modelo DINA), así como la generalización de este modelo (G-DINA) y dos modelos que se pueden derivar del anterior (A-CDM y DINO). Las diferencias básicas entre estos modelos tienen que ver con la naturaleza de la interacción entre los atributos de estudiantes e ítems, esto es, con el hecho de que se requiera o no poseer todos los atributos para poder responder correctamente a la tarea planteada en una pregunta, así como con la dependencia o independencia de los distintos atributos a la hora de contribuir al éxito en la tarea planteada a los estudiantes.

Conflict of Interest

The authors of this article declare no conflict of interest.

Note

[1]Este resumen ha sido realizado por la editora del número, María José Navas.

References

Bachman, L., & Palmer, A. (2010*). Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 42*, 75-83.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*, 163-183.

de la Torre, J. (2012). Application of the DINA Model Framework to Enhance Assessment and Learning. In M. Mok (Ed.), *Self-directed learning oriented assessments in the Asia-Pacific* (pp. 92-110). New York: Springer.

de La Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199.

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*, 450-469.

de la Torre, J., & Lee, Y. S. (2010, April). *Item-level comparison of saturated and reduced cognitive diagnosis models*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models.*Handbook of statistics,26*, 979-1030.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301-321.

Hartz, S. M. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practicality* (Unpublished doctoral dissertation).

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*, 262-277.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

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

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205-237.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment* (CRESST Report 800). Los Angeles, CA: The National Center for Research on Evaluation, Standards, and Student Testing, UCLA.

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

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives, 1*, 3-62.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*, 65-77.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment* (National Research Council's Committee on the Foundations of Assessment). Washington, DC: National Academies Press.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4-14.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Tatsuoka, C., Clements, D. H., Sarama, J., Izsak, A., Orril, C. H., de la Torre, J., & Khasanova, E. (in press). Developing workable attributes for psychometric models based on the Q-matrix. *Journal of Research on Mathematics Education.*

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods, 11*, 287-305.

Tjoe, H., & de la Torre, J. (2012). Proportional reasoning problems: Current state and a possible future direction. In S. J. Cho (Ed.), *Proceedings of the 12th International Congress on Mathematics Education* (pp. 6741-6750).

Tjoe, H., & de la Torre, J. (2013a). Designing cognitively-based proportional reasoning problems as an application of modern psychological measurement models. *Journal of Mathematics Education, 6*, 17-22.

Tjoe, H., & de la Torre, J. (2013b). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal, 26*, 237-255.

Tjoe, H., & de la Torre, J. (2014). On recognizing proportionality: Does the ability to solve missing value proportional problems presuppose the conception of proportional reasoning? *The Journal of Mathematical Behavior, 33*, 1-7.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (2013). *National Assessment of Educational Progress (NAEP), 2013 Mathematics Assessment.* Retrieved from http://nces.ed.gov/nationsreportcard/itemmaps/index.asp

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