Validity and Reliability of Situational Judgement Test Scores: A New Approach Based on Cognitive Diagnosis Models

Miguel A. Sorrel¹, Julio Olea¹, Francisco J. Abad¹, Jimmy de la Torre², David Aguado³ and Filip Lievens⁴

Abstract
Conventional methods for assessing the validity and reliability of situational judgment test (SJT) scores have proven to be inadequate. For example, factor analysis techniques typically lead to nonsensical solutions, and assumptions underlying Cronbach’s alpha coefficient are violated due to the multidimensional nature of SJTs. In the current article, we describe how cognitive diagnosis models (CDMs) provide a new approach that not only overcomes these limitations but that also offers extra advantages for scoring and better understanding SJTs. The analysis of the Q-matrix specification, model fit, and model parameter estimates provide a greater wealth of information than traditional procedures do. Our proposal is illustrated using data taken from a 23-item SJT that presents situations about student-related issues. Results show that CDMs are useful tools for scoring tests, like SJTs, in which multiple knowledge, skills, abilities, and other characteristics are required to correctly answer the items. SJT classifications were reliable and significantly related to theoretically relevant variables. We conclude that CDM might help toward the exploration of the nature of the constructs underlying SJT, one of the principal challenges in SJT research.

Keywords
situational judgment tests, cognitive diagnosis models, validity, reliability

¹Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Madrid, Spain
²Department of Educational Psychology, The State University of New Jersey, New Brunswick, NJ, USA
³Instituto de Ingeniería del Conocimiento, Universidad Autónoma de Madrid, Madrid, Spain
⁴Department of Personnel Management, Work and Organizational Psychology, Ghent University, Ghent, Belgium

Corresponding Author:
Miguel A. Sorrel, Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain.
Email: miguel.sorrel@uam.es
Situational judgment tests (SJTs) have become increasingly popular for personnel selection both in the United States and Europe (McDaniel, Morgenson, Finnegan, Campion, & Braverman, 2001; Whetzel & McDaniel, 2009). SJTs are designed to evaluate candidate judgments regarding situations encountered in the workplace (Weekley & Ployhart, 2006). Test takers are asked to evaluate each course of action either for the likelihood that they would perform the action or for the effectiveness of the action. SJTs are intended to evaluate different constructs (knowledge, skills, abilities, and other characteristics; KSAOs) related to job performance, which are different from those that are measured through cognitive ability tests or personality inventories. More specifically, a recent meta-analysis shows that SJTs intend to measure constructs that could be classified into four categories: knowledge and skills, applied social skills (e.g., leadership), basic personality tendencies (e.g., integrity), and heterogeneous composites (Christian, Edwards, & Bradley, 2010).

Despite their success, various validity and reliability issues related to SJTs have not been appropriately addressed (Christian et al., 2010; Ployhart & Weekley, 2006) because, as argued in the following, conventional methods for assessing the validity and reliability of SJT scores are based on classical test theory (CTT), which are inadequate in light of the multidimensional nature of SJT items. Therefore, this article explores the use of cognitive diagnosis models (CDMs) as a promising approach that not only overcomes these shortcomings but that also offers several advantages for scoring and better understanding SJTs.

The rest of the article is structured as follows. First, we briefly review existing validity and reliability evidence for SJT scores and in the process touch on the limitations of the existing approaches. The next section provides an introduction to CDMs. We then use an empirical example to illustrate how CDMs can be used for evaluating the validity and reliability of SJT scores and compare this approach with the traditional CTT approach. The last section discusses the advantages and the disadvantages of CDMs.

Review of SJT Literature on Reliability and Validity

Similar to any type of test, validation studies should also be conducted to provide relevant information for the interpretation and use of SJT scores. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999) specifies five “sources of evidence that might be used in evaluating a proposed interpretation of test scores for particular purposes” (p. 11). These sources of evidence are test content, consequences of testing, relations to other variables, internal structure, and response processes. In the following, we discuss to what extent these sources of evidence have been evaluated in the validation of SJT scores.

With regard to evidence based on test content, the norm in the development of SJTs is to recruit and train external “subject matter experts” (SMEs) to generate critical incidents. This information is used to develop the item stems, specify the extent to which these item situations represent the job domain, and establish the response alternatives and scoring key. Generally, once experts have made these decisions and judgments, the test is considered as more or less definitive. Furthermore, it is recognized that “there is virtually no direct investigation of the relationships linking SJTs scores and test content” (Schmitt & Chan, 2006, p. 147).

A more extensive strand of SJT studies focused on both intended and unintended consequences of SJTs score interpretation and use. Most of this research examined potential adverse impact of SJT scores, test taker perceptions toward various SJT formats, and the fake-ability of SJTs in comparison to traditional tests (for reviews, see Lievens, Peeters, & Schollaert, 2008; Whetzel & McDaniel, 2009).

Next, a voluminous stream of SJT validation studies scrutinized evidence of the relation of test scores to a relevant criterion (e.g., other constructs), their criterion-related validity with respect to
performance criteria, and their incremental validity over and above other more traditional measures (see the meta-analyses of McDaniel et al., 2001; McDaniel, Hartman, Whetzel, & Grubb, 2007). Generally, SJTs were found to have corrected validities in the mid .20s and exhibited incremental validity above and beyond traditional predictors, such as cognitive ability and personality (see also Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001; Weekley & Ployhart, 2005).

In comparison to this large body of research on the criterion-related validity of SJT scores, there is much less attention devoted to how constructs underlying SJTs are specified and examined (Arthur et al., 2014; Schmitt & Chan, 2006). The meta-analysis of Christian et al. (2010), for instance, reported that about one third of the papers published about SJTs did not indicate the construct measured, did not provide enough information about these constructs, or provided only the composite score. They concluded that “test developers and researchers often give little attention to the constructs measured” (Christian et al., 2010, p. 84). In other words, although SJTs seem to partly predict performance and enhance the criterion-related validity of traditional personality and cognitive ability test scores, the underlying reasons are not clear because little is known about the nature of the constructs measured by SJTs.

Therefore, it is widely acknowledged that more specific studies about the constructs underlying SJTs are needed (Ployhart & Ehrhart, 2003). In a recent review of personnel selection research, Ryan and Ployhart (2014) posited that among all current principal lines of research in SJTs, the exploration of the nature of the constructs is the most pressing one. Such construct-level information is pivotal because it offers several theoretical and applied advantages (Christian et al., 2010), namely, understanding deeper why some tests predict work performance better than others, comparing more clearly the effectiveness of different selection methods, reducing contamination by non–job-relevant constructs, and justifying the interpretation of the scores and their fair use.

To assess the internal structure of SJTs, one of the strategies in past research typically involved obtaining evidence via factor analytic techniques. However, the application of factor analytic techniques to SJT data almost always led to “a plethora of factors that are difficult to interpret” (Lievens et al., 2008, p. 430) as well as nonsensical factor structure solutions. Hence, it is recognized that “there has been little success in understanding what SJTs really measure” (Ployhart & Weekley, 2006, p. 346). Due to the uninterpretable factor analytic results, it has been posited that SJTs are “construct heterogeneous at the item level, because one item, for example, may target several performance dimensions” (Patterson et al., 2012, p. 853). Despite the multidimensional nature of SJTs, a single composite score is generally reported in SJT research and practice. All of these findings point to the necessity of alternative approaches for examining the internal structure (dimensionality) of SJTs and for obtaining “new insights into understanding the constructs assessed by SJTs” (Whetzel & McDaniel, 2009, p. 200).

Apart from lack of progress on how the internal structure of SJTs can be better understood, little is also known about the response processes that govern the ways in which individuals respond to SJT items. In fact, different possibilities exist regarding how individuals might respond to SJT items and solve them on the basis of their ability/skills. For instance, if a particular item includes several skills, are test takers required to master each of the skills to produce the most accurate answer (i.e., a noncompensatory model)? Or, could mastery of one of the skills compensate for the lack of mastery of the other skills (i.e., a compensatory model)? Unfortunately, these different possibilities in how individuals might respond to SJT items have not been examined with the appropriate psychometric models. As such, there exists a need for psychometric models that can provide information not only about the statistical quality of the items but also about the correspondence between the items and the targeted cognitive processes. In other words, psychometric models are needed to evaluate, among others, the appropriateness of compensatory and noncompensatory models to shed light on the item responding processes.
Finally, with respect to reliability of SJT scores, most studies have focused on internal consistency reliability (see review of Lievens et al., 2008). Generally, the internal consistency indices reported in the SJT literature are typically low. For example, a mean of .46 was obtained in some meta-analyses (e.g., Catano, Brochu, & Lamerson, 2012). These low internal consistency reliability values do not necessarily indicate poor precision of measurement. Rather, these results could reflect the fact that Cronbach’s alpha is not appropriate for assessing the reliability of multidimensional tests such as SJTs because Cronbach’s alpha requires that the construct domain be homogeneous (Schmidt & Hunter, 1996). In this context, homogeneity refers to unidimensionality (i.e., items measure a single latent construct). Given the heterogeneity of SJTs, even at the item level, researchers should look for other approaches for estimating reliability. Among other approaches, it has been proposed that test-retest reliability might be a particularly better measure for assessing the reliability of SJT scores (Lievens et al., 2008; Whetzel & McDaniel, 2009). However, “in most operational situations . . . it is impractical to obtain test-retest data” (Catano et al., 2012, p. 344). This underscores the needs to find other, more practicable approaches to estimate reliability of SJTs.

To recap, our review of research on the validity of SJT scores shows that prior research thus far has mainly focused on approaches to establishing validity evidence on the basis of test content, testing consequences, and relations to other variables. In contrast, there have been few successful attempts in providing evidence about the internal structure and response processes involved in solving SJT items. Moreover, our review of prior research highlighted the problems with using factor analytic techniques and Cronbach alpha for multidimensional tests such as SJTs. Our review also makes it clear that reliance on CTT has hampered further progress on these unexplored issues, which by nature are complex and may require more advanced psychometric models.

Thus, given these shortcomings in existing research on the validity and reliability of SJT scores, a new psychometric approach in examining the nature of constructs in SJTs is needed. Consistent with recommendations from a recent review on SJT research (Weekley, Hawkes, Guenole, & Ployhart, 2015, p. 301), we propose a specific set of latent trait measurement models, namely, cognitive diagnosis models, as an alternative psychometric approach to obtain evidence on the validity of SJT scores, assess their reliability, and score the different KSAOs that are theoretically measured by the SJT.

**Cognitive Diagnosis Models: A Tutorial**

In the past few years, there has been an increasing interest in psychometric models referred to as cognitive diagnosis models. CDMs are latent trait measurement models that explicitly allow for inferences about the underlying cognitive processes involved in responding to items and the manner in which these processes interact. In this sense, CDMs establish a link between cognitive psychology and statistical modeling. Earlier applications of CDMs are found in *cognitively diagnostic educational assessment* (Leighton & Gierl, 2007; Nichols, Chipman, & Brennan, 1995). The information that these models provide has been used for diagnosing students’ strengths and weaknesses, thereby giving teachers information that can be used to design instruction and intervention.

CDMs emerged from different fields: theory of classification (restricted latent class models; Haertel, 1989), item response theory (linear logistic test model; Fischer, 1973), and mathematical psychology (knowledge space theory; Doignon & Falmagne, 1999). Based on these different approaches, CDMs have many labels (e.g., *cognitively diagnostic models*, Henson & Douglas, 2005, *cognitive psychometric models*, Rupp, 2007; *structured IRT models*, Rupp & Mislevy, 2007).

CDMs are multidimensional, categorical latent-trait models developed primarily for assessing examinee mastery and nonmastery of a set of skills (e.g., competencies, task, knowledge, and cognitive process). Unlike traditional item response theory (IRT) models, which generally involve continuous latent variables, CDMs involve latent variables that are binary (e.g., mastery vs.
nonmastery). In the CDM literature, these categorical latent variables have been generically referred to as *attributes*. The number of attributes is denoted by $K$, and the attribute profile of respondent $i$ is denoted by $\alpha_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}\}$, where $\alpha_{ik} = 1$ or $0$ represents mastery or nonmastery of attribute $k$, respectively. CDMs are inherently confirmatory in nature as they involve a prespecified loading structure. The loading structure of a CDM, which is commonly known as Q-matrix (Tatsuoka, 1983), is a mapping structure that indicates the KSAOs required for successfully answering each individual item. A Q-matrix can be viewed as a cognitive design matrix that makes explicit the internal structure of a test. Table 1 shows the initial Q-matrix for the 23-item SJT that will be used as example in this article.

As can be seen from the table, for this test, $K = 4$ attributes are assumed to underlie the process of responding. Consider the first five items in Table 1: Items 1 and 3 require attribute 1 only; item 2 requires both attributes 2 and 3; item 4 requires attribute 4 only; and item 5 requires both attributes 1 and 2. Items 1 and 2 are shown in Figure 1. Item 1 measures study habits. Students who engage in regular acts of studying probably will answer this item correctly. Item 2 measures study attitudes and helping others. More likely than not, students who approve the broader goals of education (e.g., education should be within everyone’s reach) and tend to help others will correctly answer this item.

Confirmatory factor analysis (CFA) models and IRT models usually have a *simple structure*, that is, each item loads only on one factor (for a detailed discussion, see McDonald, 1999). Factors as defined in these models are generally broader constructs (e.g., numerical ability). In contrast, in the case of CDMs, attributes are more narrowly defined (e.g., converting a whole number to a fraction).

### Table 1. Initial Q-matrix.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: 1 = the attribute is required to choose the most effective response option; 0 = the attribute is not required to choose the most effective response option.
In addition, each item typically requires more than one attribute. This leads to a complex loading structure where each item is specified in relation to multiple attributes. This complex loading structure, in terms of multidimensional IRT, is known as within-item multidimensionality (Adams, Wilson, & Wang, 1997) and is denoted by "1s" in the Q-matrix. As noted by Schmitt and Chan (2006), SJTs tend to be multidimensional, even at the item level. Thus, in SJTs it is necessary for items to load on more than one factor. CDMs could be understood as an extension of traditional multidimensional IRT and CFA models that are particularly suitable to this kind of construct and complex loading structure.

CDMs are also called restricted (i.e., confirmatory) latent class models because the number of latent classes is restricted by the number of attributes involved in answering items of a test. With K attributes underlying performance on a given test, the respondents will be classified into $2^K$ latent classes (the number 2 indicates that there are two possible outcomes for each attribute, as in, mastery or nonmastery). A generic latent class or attribute profile can be denoted by $\alpha_l$, where the subscript index goes from $l = 1$ to $2^K$. Thus, in the aforementioned example with four attributes required to perform successfully on the test items, respondents will be classified into $2^4 = 16$ latent classes. All CDMs are expressed by $P(X_j = 1|\alpha_l)$, the conditional probability of success on item $j$ given the latent class $l$. The main output of CDM for each test taker is an estimate of the attribute profile, which gives the probability that the $i$th respondent has mastered each of the attributes. These attribute profile estimates are obtained using the expected a posteriori (EAP) method. This probability can be converted into dichotomous scores (i.e., mastery or nonmastery) by comparing them to a cut-off point (usually .50; de la Torre, Hong, & Deng, 2010; Templin & Henson, 2006). Other authors (e.g., Hartz, 2002; Jang, 2005) define an uncertainty region (e.g., between .40 and .60) within which no classifications are made, thus requiring stronger evidence before conclusions about the respondent’s state of mastery with respect to a particular attribute can be drawn.

A general CDM, called the generalized deterministic inputs, noisy “and” gate (G-DINA) model, was proposed by de la Torre (2011). The G-DINA model describes the probability of success on item $j$ in terms of the sum of the effects of involved attributes and their interactions. This model partitions

---

**Figure 1.** Items 1 and 2 of the situational judgment test (Peeters & Lievens, 2005). Most appropriate answer is shown in bold.

**ITEM 1:** When studying for an exam, do you find that you reach best results when:

a. **You start planning and setting aside time in advance**

b. work in a clean environment, even if it means taking time away from studying

c. wait for inspirations before becoming involved in most important study tasks

d. wait until the last day or so to study, knowing that you have to get it done now

**ITEM 2:** Your professor announces in class that undergraduate students are needed to help run subjects for his upcoming study. While you would not receive any formal sort of extra credit, the professor would appreciate any volunteers. Given the following choices, which option would you choose?

a. Examine your schedule and offer to volunteer a couple hours a week when it is personally convenient.

b. **Examine your schedule and offer to volunteer as many hours as you can.**

c. Realize that you would have to give up some of your free time and choose not to volunteer.

d. Offer to run subjects only if you are paid.
the latent classes into \(2^{K_j}\) latent groups, where \(K_j^*\) is the number of attributes required for item \(j\). For example, Item 2 in Figure 1 requires two of the four attributes. These two attributes lead to four latent groups: those who mastered both attributes, one of the attributes, or none of the attributes. Each latent group represents one reduced attribute vector \(\alpha_j^*\) and has an associated probability of success, written as

\[
P(X_{ij} = 1|\alpha_j^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk} \alpha_k + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*} \delta_{jk'k} \alpha_k \alpha_k' \cdots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_k'
\]

where \(\delta_{j0}\) is the intercept for item \(j\) (i.e., the probability of a correct response to an item when none of the required attributes for the item has been mastered), \(\delta_{jk}\) is the main effect due to \(\alpha_k\) (i.e., the change in the probability of a correct response as a result of mastering a single attribute), \(\delta_{jk'k}\) is the interaction effect due to \(\alpha_k\) and \(\alpha_k'\) (i.e., the change in the probability of a correct response due to the mastery of both attributes), and \(\delta_{j12\ldots K_j^*}\) is the interaction effect due to \(\alpha_1, \ldots, \alpha_{K_j^*}\) (i.e., the change in the probability of a correct response due to the mastery of all the required attributes).

The G-DINA model subsumes several commonly encountered CDMs. These include the DINO (deterministic input, noisy “or” gate; Templin & Henson, 2006) and DINA (deterministic input, noisy “and” gate; Haertel, 1989; Junker & Sijtsma, 2001) models. If several attributes are required for correctly answering the items, the DINA model can be obtained from the G-DINA model by setting to zero all terms except for \(\delta_{0}\) and \(\delta_{j12\ldots K_j^*}\); in the case of DINO model, there are also only two parameters per item, namely \(\delta_{0}\) and \(\delta_{jk}\), with the important exception that \(\delta_{jk}\) is constrained to be equal to \(\delta_{jk} = -\delta_{jk'} = \cdots = (-1)^{K_j^*+1}\delta_{j12\ldots K_j^*}\) for \(k = 1, \ldots, K_j^*\), \(k' = 1, \ldots, K_j^* - 1\), and \(k'' > k', \ldots, K_j^*\), so that some lower-order terms will be cancelled by the corresponding high-order terms. The DINA is a noncompensatory model that divides respondents in those who have mastered all measured attributes and those who are lacking at least one measured attribute, whereas the DINO is a compensatory model that divides respondents in those who master at least one measured attribute and those who are lacking all measured attributes. In this respect, the DINA model involves a conjunctive process, whereas the DINO model involves a disjunctive process. Figure 2 gives a graphical representation of an item requiring two attributes when it conforms to the DINA model, the DINO model, or the more general model (i.e., the G-DINA model).

The characteristics of CDMs discussed previously make CDM suitable for modeling the responses to a SJT. We identify four sequential steps in the application of CDMs to SJTs (see Figure 3). The first step is to develop the Q-matrix. It involves specifying the skills that are underlying performance on the SJT items and an initial Q-matrix. Next, one evaluates whether some of the original attribute specifications need to be changed on the basis of the analysis of empirical data. Once the final Q-matrix has been determined, the second step is the selection of an appropriate CDM on the basis of absolute and relative model fit. The third step consists of interpretation of the item and person parameter estimates of the selected model. Finally, the fourth step consists of searching for validity and reliability evidence of the person parameter estimates. We follow these steps in our empirical example in the following.

Assessment of SJTs Through Cognitive Diagnosis Models

This article presents a new approach to the assessment of SJTs, which aims to account for the multidimensional structure of tests. It has been shown in a prior study (García, Olea, & de la Torre, 2014) that CDMs could achieve an accurate fit to SJT data and the scores obtained could be properly interpreted. The present article substantially extends this initial work by highlighting CDMs’
Figure 2. This figure depicts the probability of correctly answering an item requiring two attributes for deterministic input, noisy “and” gate (DINA), deterministic input, noisy “or” gate (DINO), and generalized deterministic inputs, noisy “and” gate (G-DINA) models. Model parameters are denoted by $\delta$.

Figure 3. Sequential steps in the application of cognitive diagnosis models (CDMs).
usefulness in the context of reliability assessment and establishing the validity of SJTs. More specifically, this study is intended to address the following validity and reliability concerns:

1. What is the internal structure of the SJT? A CDM requires establishing a link between the items and the attributes through the Q-matrix specification. This task is typically conducted by domain experts. The recent empirical-based validation method proposed by de la Torre and Chiu (2015) then allows checking the Q-matrix generated by these experts. The Q-matrix specification and the model-fit results include information about the structural aspect, that is, how many attributes are involved at the test level, at the item level, and the relationships among them.

2. What is the general cognitive model that test takers engage in when responding to SJT items? The study of the absolute and relative fit of the different CDMs provides information about the general response processes required to solve the items. That is, we examine whether the sample of test takers engage in particular cognitive processes (e.g., conjunctive or disjunctive) when responding to the SJT.

3. Why are SJT scores good predictors of relevant theoretically relevant variables? As noted previously, SJT scores yield moderate criterion-related validity coefficients, and it is pivotal to better understand how and why SJT scores relate to the criteria and correlates. An explicit examination of the attributes measured by the SJT allows for this issue to be examined.

4. What is the reliability of the SJT assessment? As shown in the following, CDMs enable to address this question taking into account the heterogeneity of SJTs. We can use the calibrated model to generate simulate data, estimate the attribute profile for each test taker, and calculate the proportion of times that each test taker is classified correctly to the known attribute state (thus producing an estimate of attribute classification accuracy).

**Demonstration Example**

This section illustrates how CDMs can be applied to SJTs. The data for the present study were taken from the administration of an SJT composed of 23 items that present situations about various student-related issues (e.g., studying for exams and accomplishing assignments). This SJT was developed by Bess and Mullins (2002) and previously used by Peeters and Lievens (2005). By way of example, the first two SJT items are shown in Figure 1. As described in Peeters and Lievens, a total of 138 second-year psychology students from a large Belgian university participated in the study as a part of introductory courses about psychological testing and assessment. The sample was predominantly female (84%). The theoretically relevant variables (i.e., criteria and correlates) examined were grade point average (GPA, computed as the average of students’ first- and second-year GPAs), student scores on the Advances Progressive Matrices (APM; Set II; Raven, Raven, & Court, 1998), and NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992) self-report ratings (neuroticism, extroversion, openness to experience, agreeableness, and conscientiousness). Although the same data were used in Peeters and Lievens, CDM was not used in that study.

All the following analyses were carried out with the R (R Core Team, 2014) packages “CDM” (Robitzsch, Kiefer, George, & Uenlue, 2015) (functions for cognitive diagnosis modeling) and “CTT” (Willse, 2014) (a function for classical test theory analysis). The code can be easily adapted to different data sets and can be requested by contacting the corresponding author.

**Q-Matrix Development**

As pointed out by Li and Suen (2013), when developing a new Q-matrix, it is common to adopt the following procedure (Buck et al., 1998): (a) Develop an initial list of skills, (b) construct an initial Q-
matrix, (c) analyze data using an appropriate CDM with the developed Q-matrix, and (d) modify the initial Q-matrix based on statistics for each skill along with the theoretical importance of the skill. We performed our analysis according to these steps.

**Initial determination of list of skills.** Given that the attributes are an essential part of the Q-matrix, it is important to use prior research, theory, and job analytic information for determining them. Other cognitive approaches such as think-aloud protocols have been also successfully employed to gather information about the possible cognitive processes (e.g., Li & Suen, 2013). Therefore, we relied on these information sources to come up with an initial list of attributes relevant to the SJT in our empirical example. In particular, our SJT consists of 23 items that present situations about various student-related issues. In the following, we outline the concepts that could underlie this specific SJT and how they might be linked to the theoretically relevant variables.

There is now relative consensus that performance comprises of both task and contextual performance (Motowidlo, Borman, & Schmit, 1997). Task performance involves behaviors that are directly relevant to core job functions, whereas contextual performance refers to behaviors to enhance the social and psychological climate in organizations. This theoretical distinction is made not only in the job performance domain but also in the academic performance domain (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004).

Regarding dimensions underlying task performance in a student context, the meta-analysis of Créde and Kuncel (2008) revealed that study habits and study attitudes had incremental validity over standardized tests and previous grades in predicting academic performance (see also Aquino, 2011; Proctor, Prevatt, Adams, Reaser, & Petscher, 2006). Therefore, study habits and study attitudes were included in the initial list of attributes covered by the SJT in our example.

Regarding contextual performance, one of the main constructs covered is organizational citizen behavior (OCB; Bateman & Organ, 1983; Smith, Organ, & Near, 1983), which is defined by two major dimensions: (a) helping others and (b) generalized compliance (i.e., following rules and procedures; Organ, 1988). Importantly, such contextual activities are often similar across jobs and organizations (also known as transversal competences). Therefore, helping others and generalized compliance were also included in the initial list of attributes covered by the SJT in our example. Taking all of the aforementioned into account, an initial list of skills that was hypothesized to underlie this SJT was developed. Table 2 shows the four attributes (study habits, study attitudes, helping others, and generalized compliance) underlying performance on this SJT.

Next, we also put forward hypotheses related to the associations of these four attributes with theoretically relevant criteria and correlates. According to Motowidlo et al. (1997), variation in task performance is influenced by cognitive ability, whereas personality influences variation in contextual performance. Empirical findings have generally supported that personality factors predict contextual performance. In particular, three meta-analytic studies reported that conscientiousness, extraversion, neuroticism, and agreeableness are moderately correlated to cooperative contextual performance (Hough, 1992; Mount, Barrick, & Stewart, 1998; Organ & Ryan, 1995). LePine and Van Dyne (2001) found a similar pattern of results: Conscientiousness, extraversion, and agreeableness were more highly related to cooperative behavior than to task performance ($r = .17$ vs. $r = -.05$, $r = .14$ vs. $r = -.07$, and $r = .18$ vs. $r = -.03$, respectively). The correlation between neuroticism and cooperative behavior, however, was not significantly higher than the correlation between neuroticism and task performance ($r = .05$ vs. $r = .09$). Openness was related to neither task performance nor cooperative behavior ($r = -.11$ and $r = -.07$, respectively). Although there exists less research on the generalized compliance dimension, Konovsky and Organ (1996) found that it was significantly related to conscientiousness ($r = .15$).
Concerning task performance, seven meta-analysis studies demonstrated consistent relationships between conscientiousness and task performance (the $r$ coefficients vary from .20 to .31) across various occupational groups (Barrick & Mount, 1991; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Hurtz & Donovan, 2000; Salgado, 1997; Tett, Jackson, & Rothstein, 1991). Although it has been proposed that agreeableness may be an important predictor of task performance (Barrick & Mount, 1991), there is generally no evidence supporting this claim (Borman, White, & Dorsey, 1995; Hough et al., 1990; Hurtz & Donovan, 2000; Kamdar & Van Dyne, 2007; LePine & Van Dyne, 2001; Salgado, 1997).

Thus, given the aforementioned backdrop, we hypothesized that personality factors would be more highly related to the contextual performance dimensions of helping others and generalized compliance. Conversely, we hypothesized that cognitive ability and conscientiousness would be more highly related to task performance–related attributes such as study habits and study attitudes. In addition, we hypothesized that GPA would be more highly related to the studies-related attributes.

**Construction of the initial Q-matrix.** Four experts participated in an expert rating task. All of them were senior researchers with expertise in competency modeling and with extensive experience in teaching at the university level, and their native language was Spanish. The operational definitions of the four attributes were presented for their review and critique. The experts were asked to identify the
attributes needed for each item, thereby building the Q-matrix. The experts were also asked to specify the extent to which they were certain of their decisions. They employed the following system: 0 = it is certain that the attribute is not measured by the item, 1* = it is possible that the attribute is measured by the item, 1 = it is certain that the attribute is measured by the item. A Delphi process was used consisting of three rounds. In the first round, the experts were asked to identify the attributes needed for each item. In the second round, each Delphi participant was anonymously provided with the decisions of the other experts. This round provided an opportunity for participants to revise their judgments. Finally, in the third round, the four experts met in person and discussed in detail their opinions and settle the remaining differences. As done in Li and Suen (2013), we computed the Fleiss’s kappa statistic (Fleiss, 1971) to evaluate the interrater reliability of the judgments made. We considered Landis and Koch’s (1977) guidelines for interpreting kappa values, with values from .0 to .20 indicating a slight agreement, .21 to .40 a fair agreement, .41 to .60 a moderate agreement, .61 to .80 a substantial agreement, and .81 to 1 an almost perfect or perfect agreement. On the basis of the available evidence, we built the initial Q-matrix.

The experts’ ratings across the three rounds are shown in Table 3. With regard to the first round, the Fleiss’s kappa coefficients were .81 for helping others and generalized compliance and .53 for study habits indicating almost perfect and moderate agreements, respectively. However, the coefficient was only .17 for study attitudes. One possible reason for this is that this attribute is much more subjective than the other attributes, which made defining its behavioral outcomes more difficult. In the second round, when the experts were anonymously provided with the decisions made by the other experts, a high degree of agreement was achieved (the kappa coefficient for study attitudes increased up to .57). Finally, in the third round, a total agreement was achieved. The resulting attribute-item associations defined the initial Q-matrix (see Table 1). As can be seen, 11 items involved only one attribute, 8 items involved two attributes, and 4 items involved three attributes.

**Verification of the initial Q-matrix: Analysis of empirical data using an appropriate CDM.** There are many studies focused on the effect of Q-matrix misspecifications (e.g., de la Torre, 2008; Rupp & Templin, 2008a). In general, the results suggest that whenever a Q-matrix row is underspecified (i.e., a 1 is changed to a 0), the response probabilities for nonmasters of all measured attributes are overestimated (i.e., the items appear “easier”). In contrast, whenever a Q-matrix row is overspecified (i.e., a 0 is changed to a 1), we underestimate the response probabilities for masters of all measured attributes (i.e., the items appear “harder”). In addition, misspecifications in the Q-matrix may have important effects on the classification rates. Once the initial Q-matrix is specified, it is therefore important to verify its correctness. Otherwise, we cannot address any model misfit attributable to the Q-matrix.

To accomplish this, we used the test takers’ responses to the SJT to empirically validate the Q-matrix following the general method of empirical Q-matrix validation recently proposed by de la Torre and Chiu (2015). This method is based on a discrimination index, which can be used in conjunction with the G-DINA model. Thus, the proposed index does not require making an assumption about which specific models are involved. The general discrimination index is defined as

\[
\mathcal{D}_j^2 = \sum_{i=1}^{K_j} w(\alpha_{ij}) \left[ P(\alpha_{ij}) - \bar{P}_j \right]^2,
\]

where \( w(\alpha_{ij}) \) is the probability of the reduced attribute pattern, \( \alpha_{ij} \), \( P(\alpha_{ij}) \) is the probability of success of the reduced attribute pattern \( \alpha_{ij} \), and \( \bar{P}_j = \sum_{i=1}^{K_j} w(\alpha_{ij}) P(\alpha_{ij}) \) is the mean success probability. This discrimination index measures the extent to which an item can differentiate between the
different reduced attributed vectors based on their success probabilities and is minimum (i.e., equal to zero) when \( P(\alpha_j/C) = \ldots = P(\alpha_j/C_K) = \bar{P}. \) The maximum value of \( \zeta_2^2 \) for item \( j \) (i.e., \( \zeta_{j_{max}}^2 \)) is obtained when all attributes are specified (de la Torre & Chiu, 2015). In addition, de la Torre and Chiu (2015) define the proportion of variance accounted for (PVAF) by a particular q-vector relative to this maximum as \( B_2^2 = B_{2_{max}}^2. \)

Modification of the initial Q-matrix. As de la Torre and Chiu (2015) acknowledged, in many applied situations, Q-matrix recommendations based on the empirical validation procedure method can differ, sometimes markedly, from the Q-matrix based on expert opinions. In our case, changes suggested by the empirical validation were implemented if the following criteria were fulfilled: (a) gains in terms of the \( \zeta_2^2 \) (i.e., \( \Delta PVAF \)) were considered substantial (i.e., at least .30) and (b) changes made theoretical sense. To explore whether the changes suggested had theoretical basis, we took into consideration the ratings across the three rounds of the expert task (see Table 3). Note that experts were allowed to express uncertainty about their ratings (noted with * in Table 3). At this step, a suggested change was determined to have theoretical basis when at least one expert identified with certainty that the attribute as necessary/unnecessary. Finally, for the changes that met the criteria, we assessed the model fits with the Akaike Information Criterion (AIC; Akaike, 1974) to determine the final Q-matrix.

Although many of the suggested changes led to an improvement in the item discrimination, only Items 2 and 17 were found to also have some theoretical basis. For example, Item 2 in Figure 1 originally required attributes 2 and 3. As shown in Table 4, the suggested attribute specification prescribed all the attributes with \( \Delta PVAF = .71 \). However, the experts recommended only attribute 1, but not attribute 4, with certainty (see Table 3, Round 1). This change has an associate \( \Delta PVAF = .60 \). The same was true for item 17.

### Table 3. Expert Ratings for the Items of the Situational Judgment Test.

<table>
<thead>
<tr>
<th>Item</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1, 2*, 3*</td>
<td>2, 3</td>
<td>2, 3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>1, 4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1, 2*</td>
<td>1, 2</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>1*</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>1, 2*, 4</td>
<td>1, 2, 4</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>1*</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>2*, 3*</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>1*, 2*</td>
<td>1, 2</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>1*, 3*</td>
<td>1, 4</td>
<td>1, 4</td>
</tr>
<tr>
<td>13</td>
<td>1, 2*, 4*</td>
<td>1, 4</td>
<td>1, 4</td>
</tr>
</tbody>
</table>

**Note:** Attributes in bold were considered necessary by the four experts. Attributes: 1 = study habits; 2 = study attitudes; 3 = helping others; 4 = generalized compliance.

*Three experts considered the attribute necessary. **Two experts considered the attribute necessary. *One expert considered the attribute necessary.

*At least one expert expressed uncertainty about the necessity of the attribute.
A change in Item 2 only, a change in Item 17 only, and changes in both Items 2 and 17. Based on the AIC, the best results were obtained for changing only the specification for Item 2. Therefore, we modified only the attribute specification for Item 2.

### Selection of the Appropriate CDM

Each of the CDMs described in the introduction section specify the relationships among the postulated attributes in a different way. Whereas the DINA and DINO are conjunctive and disjunctive models, respectively, the G-DINA model is a general model that allows for both types of relationships within the same test. To select the most appropriate CDM for the test, one can assess the absolute and relative fit of each model. Considering that the DINA and DINO models are nested in the G-DINA model (de la Torre, 2011), one can employ the likelihood ratio (LR) test to evaluate their relative fit. The DINA and DINO models will always have a lower log-likelihood given that they are specific cases of the G-DINA model, but it is necessary to test whether the observed difference in model fit is statistically significant. The LR test does this by comparing the log-likelihoods of the models. This statistic is widely employed in other statistical models (e.g., structural equation models) for comparing nested models. It is assumed to be asymptotically $\chi^2$ distributed with degrees of freedom equal to the difference between the numbers of parameters of the general and the reduced models. If the LR is significantly different from 0, the general model fits the data significantly better than the reduced model. Regarding absolute fit, we evaluated how well each proposed model reproduces the observed data. This is typically done by assessing indices based on residual analysis. We evaluated item fit statistics on the basis of the standardized residuals between the observed and predicted Fisher-transformed correlations of item pairs (Chen, de la Torre, & Zhang, 2013). To evaluate the absolute fit, Chen et al. (2013) proposed examining the z-score of the maximum absolute residual. If the evaluated model fits the data, this statistic should not be significantly different from zero. This approach is analogous to the inspection of the residual correlation matrix in structural equation modeling.

Table 5 shows the indices calculated for test fit and item fit for the G-DINA, DINA, and DINO models. The two $\chi^2$ tests, each one with 44 degrees of freedom, corresponding to the likelihood ratio tests resulting from comparing the G-DINA model with the DINA (LR = 85.06) and DINO (LR = 82.55) models, were both significant ($p < .05$). These results indicate that the more parsimonious models led to a significant loss of fit. Absolute item fit statistics also indicated that the G-DINA model had better fit than the reduced models. When the G-DINA is fitted to the data, the z-score of the maximum absolute Fisher-transformed was not significant at $\alpha$-level of .05 after applying the Holm-Bonferroni correction (Holm, 1979). Based on the previous information, the DINO and DINA model were discarded, and the G-DINA model was further examined for its adequacy to model the SJT data.

### Table 4. Largest $\chi^2$ and PVAF of Item 2 for Different Numbers of Attribute Specifications.

<table>
<thead>
<tr>
<th>Item</th>
<th>Attribute Specification</th>
<th>$\chi^2$</th>
<th>PVAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1000</td>
<td>0.05</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td>0110&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.02</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>1110</td>
<td>0.06</td>
<td>.86</td>
</tr>
<tr>
<td></td>
<td>1111&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.07</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: $\chi^2$ = general discrimination index; PVAF = proportion of variance accounted by the q-vector relative to the $\chi^2_{max}$.  
<sup>a</sup>Original.  
<sup>b</sup>Suggested.
Table 5. Model Fit Indices for Different Cognitive Diagnosis Models.

<table>
<thead>
<tr>
<th>Model</th>
<th>loglike</th>
<th>Npars</th>
<th>LR</th>
<th>df</th>
<th>p Value</th>
<th>abs(fcor)</th>
<th>z-Score</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-DINA</td>
<td>-1,822.15</td>
<td>101</td>
<td></td>
<td></td>
<td></td>
<td>.28</td>
<td>3.28</td>
<td>.13</td>
</tr>
<tr>
<td>DINA</td>
<td>-1,864.68</td>
<td>57</td>
<td>85.06^a</td>
<td>44</td>
<td>&lt;.001</td>
<td>.32</td>
<td>3.75</td>
<td>.02</td>
</tr>
<tr>
<td>DINO</td>
<td>-1,863.43</td>
<td>57</td>
<td>82.55^b</td>
<td>44</td>
<td>&lt;.001</td>
<td>.32</td>
<td>3.71</td>
<td>.03</td>
</tr>
</tbody>
</table>

Note: loglike = log likelihood; Npars = number of model parameters; LR = likelihood ratio; abs(fcor) = maximum absolute Fisher-transformed correlation; DINA = deterministic input, noisy “and” gate; DINO = deterministic input, noisy “or” gate; G-DINA = generalized deterministic inputs, noisy “and” gate.

^aG-DINA versus DINA. ^bG-DINA versus DINO.

**Interpretation of Model Parameter Estimates**

**Item parameter estimates.** In the next step, we described the items using both CTT and CDM indices. Regarding CTT indices, we used the proportion correct or item difficulty (\(P\)) and corrected point-biserial correlation (\(r_{pb}\)). Based on the item parameter estimates for the selected CDM (G-DINA), \(\hat{\zeta}^2\) was computed. We also examined the difference between the probabilities of success for individuals who mastered none (i.e., \(P(0)\)) and all of the attributes required (i.e., \(P(1)\)). For example, if item \(j\) measures \(K^*_j = 2\) attributes, this difference is computed as \(P(1) - P(0)\). Unlike \(\hat{\zeta}^2\), this difference can be negative.

Table 6 presents the estimates of \(P_j\), \(r_{pb}\), G-DINA parameters, \(P(1^*_j) - P(0^*_j)\), and \(\hat{\zeta}^2\). In general, for the G-DINA model, good items are those that have small baseline probability (i.e., \(P(0^*_j)\)) and the probability of getting a correct response increases as the number of mastered attributes increases. For example, in the case of Item 5, the probability that respondent \(i\) with latent class \(\alpha_i\) will correctly answer the item, an indicator for attributes 1 and 2, can be written as follows:

\[
P(X_{i5} = 1|\alpha_i) = \delta_{50} + \delta_{51}\alpha_{i1} + \delta_{52}\alpha_{i2} + \delta_{512}\alpha_{i1}\alpha_{i2} = .62 + .07\alpha_{i1} + .27\alpha_{i2} + .04\alpha_{i1}\alpha_{i2}
\]

Thus, the baseline probability is rather high (\(\delta_{50} = P(00) = .62\)). The increment in the probability of correctly answering the item as a result of the presence of \(\alpha_1\) is small (\(\delta_{51} = P(10) - P(00) = .09 - .62 = .07\)), whereas mastering \(\alpha_2\) increases the probability of correctly answering the item up to .89 (\(P(01) = \delta_{50} + \delta_{52} = .62 + .27 = .89\)). The probability of success for respondents mastering both attributes is approximately 1 (\(P(11) = \delta_{50} + \delta_{51} + \delta_{52} + \delta_{512} = .62 + .07 + .27 + .04 = 1\)). The interaction effect due to the presence of both attributes is low (\(\delta_{512} = P(11) - P(00) - P(10) = 1 - .62 - .07 - .27 = .04\)).

As can be seen from Table 6, some of the items with the lowest \(\hat{\zeta}^2\) had some of the highest \(P(0^*_j)\). For example, Item 13 was one of the least informative because nonmasters of the required attributes (1 and 4) have a substantial chance of guessing the correct answer, \(P(00) = .75\). Indeed, it was found that a high percentage of the respondents answered the item correctly (\(P_{13} = .91\)).

To further explore the relationships between the G-DINA and CTT indices, the correlation between these indices was computed (see Table 7). We found a highly significantly positive correlation between \(P_j\) and \(P(0^*_j)\) and \(P(1^*_j)\); the CTT discrimination index, \(r_{pb}\), was highly correlated with \(P(1^*_j) - P(0^*_j)\) and moderately correlated to \(\hat{\zeta}^2\) and \(P(0^*_j)\). The two item discrimination indices in CDM were highly correlated.
Person parameter estimates. Table 8 shows the attribute class probabilities and the class expected frequency in the sample of 138 respondents. The second column shows the possible attribute profiles for all the 16 latent classes. As the third column shows, the attribute profile of $\alpha_{16} = \{1111\}$ had the highest class probability of about .32. That is, approximately 32% of the respondents (as shown in the fourth column, 44 respondents) were classified as belonging to this latent class and therefore
were expected to master all of the four attributes. After applying the cut-off points (i.e., >.60 for mastery and <.40 for nonmastery), the percentage of examinees who did not receive a classification was 1%, 4%, 7%, and 2% for attributes 1, 2, 3, and 4, respectively.

Figure 4 depicts an example of how CDMs allow for a finer-grained analysis of the test takers’ strengths and weaknesses. Test takers with the response pattern A correctly answered 9 items correctly. If we look at the Q matrix depicted in Table 1, we notice that these test takers correctly answer 4 out of the 6 items measuring generalized compliance (attribute 4). Thus, we estimate that they have a high probability (91%) of mastering this attribute. On this basis, these test takers are classified as masters of generalized compliance. Test takers with the response pattern B correctly answered 14 items correctly. We estimate that they have a high probability of mastering attributes 1, 2, and 4 (76%, 76%, and 93%, respectively). Note that despite the fact that these test takers fail at 6 out of the 10 items measuring study habits (attribute 2), some of the items that they correctly answered are highly discriminating (e.g., Items 5, 11, and 22). This explains why these test takers were estimated to have a high probability of mastering the attribute. The most uncertain estimate of an attribute mastery probability is at .50. For this reason, we recommend employing the discussed cut-off points (i.e., .40 and .60). Thus, no classification is made for helping others (attribute 3) for test takers with the response pattern B.

**Validity and Reliability Evidences**

**Relationships among attributes and criterion/correlates.** Once the person parameter estimates were estimated (i.e., the expected probability of mastering each attribute), we computed the correlations among the attribute scores, the SJT sum score, and the criterion/correlates. To eliminate the floor and ceiling effects inherent in the attribute probabilities, we used the logit transformation. As shown in Table 9, study habits (attribute 1) was highly correlated with GPA ($r = .35$) and conscientiousness ($r = .53$), and these correlation coefficients were somewhat higher than those estimates for the SJT sum score (.30 and .46, respectively). Thus, most of the predictive power of the SJT scores is due to this single attribute. Conversely, as we hypothesized, helping others (attribute 3) was generally related to the personality measures. The pattern of correlations is similar to the one obtained for the SJT sum score. Study habits and study attitudes (attributes 1 and 2) were also related to some of

<table>
<thead>
<tr>
<th>Latent Class</th>
<th>Attribute Profile</th>
<th>Class Probability</th>
<th>Class Expected Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0000</td>
<td>.12</td>
<td>17.07</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0100</td>
<td>.01</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>1100</td>
<td>.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>0010</td>
<td>.02</td>
<td>2.73</td>
</tr>
<tr>
<td>6</td>
<td>1010</td>
<td>.05</td>
<td>6.46</td>
</tr>
<tr>
<td>7</td>
<td>0110</td>
<td>.02</td>
<td>2.99</td>
</tr>
<tr>
<td>8</td>
<td>1110</td>
<td>.10</td>
<td>14.48</td>
</tr>
<tr>
<td>9</td>
<td>0001</td>
<td>.12</td>
<td>16.69</td>
</tr>
<tr>
<td>10</td>
<td>1001</td>
<td>.04</td>
<td>5.55</td>
</tr>
<tr>
<td>11</td>
<td>0101</td>
<td>.11</td>
<td>15.14</td>
</tr>
<tr>
<td>12</td>
<td>1101</td>
<td>.07</td>
<td>10.31</td>
</tr>
<tr>
<td>13</td>
<td>0011</td>
<td>.00</td>
<td>0.00</td>
</tr>
<tr>
<td>14</td>
<td>1011</td>
<td>0.01</td>
<td>1.11</td>
</tr>
<tr>
<td>15</td>
<td>0111</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>16</td>
<td>1111</td>
<td>0.32</td>
<td>44.61</td>
</tr>
</tbody>
</table>
these personality measures. Generalized compliance (attribute 4) was not significantly related to any of the theoretically relevant variables. Although most attributes were highly intercorrelated, this was also not the case for generalized compliance (attribute 4). This attribute was not significantly related to helping others (attribute 3), and the correlations with the other attributes were moderate in size. Finally, note that neither the SJT sum score nor the attributes were significantly related to the RAVEN score (which might be due to the range restricted nature of the university student sample; see Peeters & Lievens, 2005).

Reliability of the assessment. The alpha reliability coefficient depends on the assumption that all the items reflect a single construct (Miller, 1995). Given that SJT items are typically heterogeneous, coefficient alpha can be expected to be an inaccurate measure of the true reliability (see Catano...
et al., 2012). Indeed, the internal consistency of the SJT scores (.57) in this sample was rather low. As noted previously, it therefore makes sense to use a reliability coefficient that takes into consideration the multidimensional nature of the SJT items. More importantly, from the CTT, we cannot estimate the reliability for the underlying dimensions that are being measured by the SJT. CDMs represent a new approach for assessing the reliability of these scores. A common indicator of reliability in CDM is called attribute classification accuracy, which indicates how accurately a CDM classifies test takers into correct attribute profiles.

To estimate attribute classification accuracy, we use the calibrated model to generate simulated data so that we could study the attribute classification accuracy once the true classifications are known. For this purpose, the responses from 1,600 examinees were simulated, that is 100 examinees for each one of the $2^4 = 16$ possible attribute profiles (i.e., latent classes). The model employed was the G-DINA model, and the values of the item parameters were those estimated from the empirical data. Then we fitted the G-DINA model to the simulated data set. The following cut-off points were applied to the EAP estimates: We define mastery as a posterior probability of mastering the attribute above .50 and nonmastery as a probability between below .50. We calculated the proportion of times that a test taker is classified correctly according to the true classifications. This analysis allowed determining: (a) the attribute level classification accuracy, that is, the proportion of correct classifications for each of the four attributes, and (b) the pattern level classification accuracy, which is defined as the proportion of times that a test taker is correctly classified in all the assigned attributes.

Results of this simulation study show that the attribute level classification accuracy was considerably high. The proportion of correctly classified individual attributes was always at least .85 (.95, .93, .85, and .93 for attributes 1, 2, 3, and 4, respectively). With regard to the pattern level classification accuracy, the proportion of times all the classified attributes were classified correctly was also considerably high (76%). Regarding the proportion of times that a test taker was correctly classified at least in 2 or 3 attributes, the proportions increased to .94 and .97, respectively.

Discussion

Contributions of Cognitive Diagnosis Models

To date, in the SJT domain, some of the sources of validity (those based on internal structure and response processes) and reliability have not been appropriately addressed. Therefore, it has been reiterated that the constructs SJTs measure are unknown (e.g., Christian et al., 2010; Ployhart & Weekley, 2006). This article posited that the absence of an appropriate psychometric approach has been a major obstacle to move the field forward because traditional psychometric procedures (e.g., factor analysis and Cronbach’s alpha) cannot deal with the item multidimensionality in SJTs.

In this study, we explored how the CDM approach can offer useful solutions to these predicaments. We illustrated how common validity and reliability concerns in SJT research can be addressed by assessing the Q-matrix specification, the model fit, and the item and examinee parameter estimates. As summarized in the following, we demonstrated that the advantages of CDM over CTT in providing a greater wealth of information in analyzing SJTs are fourfold.

First, we showed that the application of a CDM model allows getting a better understanding of the underlying internal structure of the SJT. In our empirical example, successful completion of the SJT was found to require four attributes: study habits, study attitudes, helping others, and generalized compliance. As we have seen, all of these attributes are positively correlated, except helping others and generalized compliance. Importantly, the empirical validation of the Q-matrix allows for the experts’ decisions and judgments to be verified. This empirical validation of the Q-matrix resulted in a new specification for one item that was supported by substantive theory as well as increased the
item’s discrimination power. On the basis of increased insight in the underlying multidimensional structure of the SJT, CDMs allow for separately scoring the different attributes that are measured by the test, which is not possible with the typical use of a single overall score in SJTs.

Second, CDMs can illuminate response processes underlying SJTs because they show which set of KSAOs are required for solving SJT items and whether or not one KSAO can potentially compensate for the others. Through the study of the model fit, we were able to determine that the G-DINA model achieved the best fit to the data, and constraining the model to be conjunctive or disjunctive (i.e., using the DINA and DINO models) led to a significant loss of fit. According to the item parameters, different types of processes were involved within the same test.3 In the case of some items (e.g., Item 23), only test takers who have mastered all the required attributes had a high probability of selecting the most effective answer. In the case of other items (e.g., Item 8), the mastery of one or more attributes could make up for lack of mastery in other attributes. There were still other items (e.g., Item 5) in which mastering each of the attributes led to an increase in the probability of success on a certain item, whereas the effect of the interaction among the attributes was negligible.

Third, we showed how CDM can provide information about the relationships of the four underlying dimensions (attributes in CDM language) in the SJT and theoretically relevant variables. As expected, student-related attributes (study habits and attitudes) were significantly related to GPA (Aquino, 2011) and conscientiousness (Barrick & Mount, 1991; Hough et al., 1990; Hurtz & Donovan, 2000; Salgado, 1997; Tett et al., 1991), and the helping others attribute was significantly related to personality (Hough, 1992; LePine & Van Dyne, 2001; Mount et al., 1998; Organ & Ryan, 1995). In this way, when we model the multidimensional nature of SJT, we gain insights into the relationships among the SJT scores and theoretically relevant variables. This also signals which attributes do not function as expected, which might trigger efforts to redesign the test at hand. Contrary to prior research (Konovsky & Organ, 1996), for instance, generalized compliance was not significantly related to any of the variables. We tentatively attribute this result to a poor representation of the construct domain of generalized compliance. There were only six items measuring this attribute, and inspection of their item content revealed that all of them represented situations in which students had to follow the norms proposed by their teacher (e.g., stick with the existing timetable). Other aspects of the generalized compliance construct such as punctuality and not wasting time were not represented in the current items.

Fourth, we illustrated how CDMS can allow for the reliability of SJT scores to be studied from an angle different from how it is traditionally done (i.e., based on Cronbach’s alpha or test-retest procedures). Test precision in CDM is similar to the logic underlying CTT. In many testing contexts, it is necessary to classify respondents into performance categories. Decision accuracy refers to the extent to which classifications based on the observed scores agree with the classifications based on the true scores. Similarly, classification accuracy in CDM is intended to measure the degree to which classifications based on observed scores matched the true attribute profile. In our empirical example, the agreement-rate calculation between true and estimated attribute profiles based on the simulated data indicated that the proportion of times that the entire attribute profile is recovered was considerably high. In addition, CDM results provided information about individual attribute classification accuracy. This enables researchers to determine whether any of the attributes was measured with low reliability. Taking the items with a high discrimination index as an example, additional assessment tasks could be designed, specifically for attributes with lower accuracy classification rates, so that the resulting SJT might achieve higher levels of reliability. These new items can be added to the calibrated item pool through linking designs, as it is often done in IRT. In the most common scenario, a group of examinees will take a set of old (i.e., calibrated) items and a set of new (i.e., uncalibrated) items.
Finally, apart from the fourfold information that test users and designers could get, CDMs also provide finer-grained information about test takers’ strengths and weaknesses. This information could be fruitfully used by HR practitioners in SJT applications, such as personnel selection and needs analyses in training programs (Weekley et al., 2015). A generic example of the prototypical feedback was shown in the empirical example. That is, the feedback consists of a list of attributes and indicates per attribute the probability that the test taker has mastered the attribute. Providing this feedback to test takers is relatively straightforward. The main point to consider when making a decision on which cut-off point to employ to convert these probabilities into profiles is the goal of the assessment (e.g., the willingness to report low-reliable profiles). If all respondents must be classified one way or another, one can employ .50 as cut-off score. On the other hand, in some applied contexts, one might be more interested in selecting high-performing (e.g., personnel selection) or low-performing (e.g., educational assessment) individuals. If that is the case, one needs to ensure that those specific patterns are accurately estimated. In addition, cognitive diagnosis computer adaptive assessments (CD-CAT) serve as one possible solution for the problem of having nonclassified individuals (for an overview, see e.g., Huebner, 2010). The termination criterion is generally based on the accuracy with which the respondents are assessed. Thus, for example, the diagnostic assessment can only be terminated when the posterior probability that a respondent belongs to a given state (i.e., mastery or nonmastery) achieves an acceptable value (e.g., less than .20 or greater than .80).

**Caveats Related to Cognitive Diagnosis Models**

Some caveats related to CDM should be acknowledged. First, we want to emphasize that the initial list of attributes should be carefully developed. As noted, this can be done via a variety of methods such as prior research, theory, job analytic information, and think-aloud protocols. It is equally pivotal to verify the Q-matrix developed (de la Torr, & Chiu, 2015), as we did in our empirical example, to correct possible misspecifications in the original Q-matrix. De la Torre and Chiu (2015) showed that the empirical validation procedure can accurately identify and correct misspecified q-entries without altering correct entries, particularly when high-quality items are involved. This is typically the case in educational assessment where items tend to be highly discriminating, but the results cannot be directly extrapolated in the case of poor-quality items. Thus, we stress the importance of relying on the expert ratings to examine these discrepancies. We also suggest doing a cross-validation in another sample to avoid the possibility of capitalization on chance, which might bias the statistical estimates.

Second, the relations between CDM and CTT deserve attention. There are various points in common between these two approaches. Lee, de la Torre, and Park (2011) explored the relationships between CDM, CTT, and IRT indices. The pattern of correlations among CTT and CDM indices that they reported is very similar to the one we obtained: Difficulty and discrimination CTT and CDM indices are typically highly correlated. We do not see this similarity in results as a limitation of CDM. Rather, it is a positive point that specific CDM indices correspond to the results of CTT indices. Our results indicate that items can provide diagnostic information (e.g., help differentiate between respondents who have mastered more attributes and respondents who have mastered fewer attributes) even if they are not developed under a CDM framework. The CTT discrimination indices may provide guidance on the diagnostic value of an item. In this way, items with low corrected point-biserial correlation can be expected to have low discrimination in CDM. In addition, as shown in our article, CDM indices provide a host of extra information over and above CTT indices. One difference between CDMs and CTT, which is a potential disadvantage of CDMs, is that their parameters must be estimated. Standard error of model parameters can be used as a measure of the precision of the estimate. Standard error estimates depend on the sample size: As sample size
increases, the standard error decreases. Note, however, that it has been shown that when the model fits the data, the DINA model parameters are invariant (de la Torre & Lee, 2010). Thus, no matter what sample of respondents takes the test, the item parameter estimates will generally be the same. This means that item parameter estimates have to be estimated only once, provided the sample is representative of the population.

A third caveat related to the application presented in the current study is that the specification of Q-matrix was done after the test was developed. This approach, referred to as retrofitting, is actually commonly found in the CDM literature. A good example is the study of Templin and Henson (2006), who demonstrated how the hypothesized underlying factors contributing to pathological gambling can be measured with the DINO model. However, in those applications, where CDM have been retrofitted to assessments constructed using a unidimensional or CTT framework, convergence problems may occur, as well as poor item, respondent, or model fit (Rupp & Templin, 2008b). Thus, a more optimal approach is to design a test from the beginning and apply these theory-based specifications during the test development process itself (de la Torre, Tjoe, Rhoads, & Lam, 2010).

Conclusion

This study proposed and illustrated how CDM can be used to explore the nature of the constructs that SJTs measure, which is one of the current and principal challenges in SJT research (Ryan & Ployhart, 2014; Weekley et al., 2015). Overall, we conclude that CDMs include a greater wealth of information in analyzing SJTs than traditional procedures based on CTT do. That is, CDM holds promise in evaluating the internal structure of the SJT, providing information about the cognitive processes underlying the responses in the SJT, clarifying how and why the SJT scores relate to other variables, and leading to a more appropriate estimation of the reliability of these scores.

Acknowledgements

The authors wish to thank associate editor Adam Meade and three anonymous reviewers for their valuable comments and suggestions on earlier versions of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by Grant PSI2013-44300-P (Ministerio de Economía y Competitividad and European Social Fund).

Notes

1. Based on the probabilities of being classified into an attribute profile given the data (i.e., $P(\alpha|X)$), the individual attribute profile can be deduced via three methods: maximum likelihood estimation (MLE), maximum a posteriori (MAP) estimation, and expected a posteriori (EAP) estimation. For a comparison among MLE, MAP, and EAP classification methods, see Huebner and Wang (2011).

2. Currently there are different programs available for estimating cognitive diagnosis models (CDMs), for example the G-DINA framework in Ox (Doornik, 2002) by de la Torre, the MDLTM program by von Davier (2005), the LCDM framework in SAS (SAS Institute Inc., 2007) and Mplus (Muthén & Muthén, 2012) by Templin, Henson, Douglas, and Homan. The main advantage of R is that it is freely available and very flexible.
3. When referring to a particular underlying latent structure and the response processes implied, it should be acknowledged that between-subjects conclusions should not be interpreted at the individual level (Borsboom, Mellenbergh, & van Heerden, 2003). Recently, this issue has been considered in measurement equivalence (Tay, Meade, & Cao, 2015).

References


Author Biographies
Miguel A. Sorrel is a PhD candidate at the Department of Social Psychology and Methodology at Universidad Autónoma de Madrid. His research interests include item response theory, cognitive diagnosis modeling, and computerized adaptive testing.
Julio Olea is a professor of methodology of behavioural sciences at Universidad Autónoma de Madrid. His teaching and research work has focused on the field of psychometric: implementation of scaling methods, cognitive diagnosis modeling, and computerized adaptive testing.
Francisco J. Abad is an associate professor at Universidad Autónoma de Madrid. He has developed his teaching and research work in the field of psychometrics: polytomous IRT models, goodness of fit, software development, and computerized adaptive testing.
Jimmy de la Torre is a professor of educational psychology at Rutgers University. His primary research interests are in the field of psychological and educational testing and measurement, particularly in the areas of item response theory, cognitive diagnosis modeling, and psychometric models for noncognitive test data.
David Aguado is an associate professor at the Universidad Autónoma de Madrid. He is the director of innovation in talent management at the Instituto de Ingeniería del Conocimiento (IIC). He has developed his teaching and research work in competency development and performance management.
Filip Lievens is full professor at the Department of Personnel Management, Work and Organizational Psychology at Ghent University. His research has influenced a variety of applied measurement issues in I-O psychology. He has published in the Annual Review of Psychology, Journal of Applied Psychology, Personnel Psychology, and Journal of Management.