GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2024) 36(2) 145-153

Psicothema

https://www.psicothema.com/es • ISSN 0214–9915 • eISSN 1886-144X

Colegio Oficial de Psicología del Principado de Asturias

Psicothema

Article

Enhancing Content Validity Assessment With Item Response Theory Modeling

Rodrigo Schames Kreitchmann[1], Pablo Nájera[2], Susana Sanz[2] and Miguel Ángel Sorrel[2]

1 Universidad Nacional de Educación a Distancia (Spain)
2 Universidad Autónoma de Madrid (Spain)

ARTICLE INFO

Received: May 24, 2023
Accepted: September 27, 2023

Keywords:
Content validity
Subject matter experts
Item response theory
Validity
Test development

ABSTRACT

Background: Ensuring the validity of assessments requires a thorough examination of the test content. Subject matter experts (SMEs) are commonly employed to evaluate the relevance, representativeness, and appropriateness of the items. This article proposes incorporating item response theory (IRT) into model assessments conducted by SMEs. Using IRT allows for the estimation of discrimination and threshold parameters for each SME, providing evidence of their performance in differentiating relevant from irrelevant items, thus facilitating the detection of suboptimal SME performance while improving item relevance scores. **Method:** Use of IRT was compared to traditional validity indices (content validity index and Aiken's V) in the evaluation of *conscientiousness* items. The aim was to assess the SMEs' accuracy in identifying whether items were designed to measure conscientiousness or not, and predicting their factor loadings. **Results:** The IRT-based scores effectively identified conscientiousness items (R^2 = 0.57) and accurately predicted their factor loadings (R^2 = 0.45). These scores demonstrated incremental validity, explaining 11% more variance than Aiken's V and up to 17% more than the content validity index. **Conclusions:** Modeling SME assessments with IRT improves item alignment and provides better predictions of factor loadings, enabling improvement of the content validity of measurement instruments.

Mejorando la Evaluación de la Validez del Contenido a Través del Modelado de la Teoría de Respuesta al Ítem

Palabras clave:
Validez de contenido
Expertos en la materia
Teoría de respuesta al ítem
Validez
Desarrollo de tests

RESUMEN

Antecedentes: Garantizar la validez de evaluaciones requiere un examen exhaustivo del contenido de una prueba. Es común emplear expertos en la materia (EM) para evaluar la relevancia, representatividad y adecuación de los ítems. Este artículo propone integrar la teoría de respuesta al ítem (TRI) en las evaluaciones hechas por EM. La TRI ofrece parámetros de discriminación y umbral de los EM, evidenciando su desempeño al diferenciar ítems relevantes/irrelevantes, detectando desempeños subóptimos, mejorando también la estimación de la relevancia de los ítems. **Método:** Se comparó el uso de la TRI frente a índices tradicionales (índice de validez de contenido y V de Aiken) en ítems de *responsabilidad*. Se evaluó la precisión de los EM al discriminar si los ítems medían responsabilidad o no, y si sus evaluaciones permitían predecir los pesos factoriales de los ítems. **Resultados:** Las puntuaciones de TRI identificaron bien los ítems de responsabilidad (R^2 = 0,57) y predijeron sus cargas factoriales (R^2 = 0,45). Además, mostraron validez incremental, explicando entre 11% y 17% más de varianza que los índices tradicionales. **Conclusiones:** La TRI en las evaluaciones de los EM mejora la alineación de ítems y predice mejor los pesos factoriales, mejorando validez del contenido de los instrumentos.

Cite as: Kreitchmann, R. S., Nájera, P., Sanz, S., & Sorrel, M. A. (2024). Enhancing content validity assessment with Item Response Theory modeling. *Psicothema, 36*(2), 145-153. https://doi.org/10.7334/psicothema2023.208
Corresponding author: Rodrigo Schames Kreitchmann, rschames@psi.uned.es

Test developers often rely on judgements made by subject matter experts (SMEs) as sources of evidence around assessment validity. In educational settings, experts may be consulted to evaluate the *alignment* of the items to the curricular standards and learning objectives, indicating whether an item is appropriate for assessing a given grade group, or has the desired depth of knowledge or cognitive complexity (e.g., Bhola et al., 2003; Webb, 2007). Similarly, in diagnostic assessments, experts are often consulted to determine the skills or abilities required for correctly solving each problem in a test (García et al., 2014; Nájera et al., 2021; Tatsuoka, 1983). Likewise, for the assessment of non-cognitive domains, such as personality, motivation or leadership, SMEs are relied upon to evaluate the appropriateness of the items regarding various aspects. For instance, they may be asked to judge if different aspects of the constructs are properly represented in a test (e.g., Polit & Beck, 2006). In a similar fashion, SMEs may assess the degree to which each item is relevant to the construct, as well as the appropriateness of other aspects related to item wording, such as clarity, lack of ambiguity, and technical quality (Fitzpatrick, 1983; Mastaglia et al., 2003; Penfield & Giacobbi, 2004). From a broad perspective, the efforts in ensuring the adequacy of the test content for the measured constructs partially constitute what is commonly called *content validity*, or, more correctly, *validity evidence based on test content* (American Educational Research Association [AERA] et al., 2014).

Gathering validity evidence for test content involves determining the test's ability to accurately measure its intended purpose. Content-based validity evidence comprises four main aspects: *domain definition, domain representation, domain relevance,* and *appropriateness of test construction procedures* (Sireci, 1998b). While the first refers to the definition of the measured domains, the others concern the test itself. Evaluating domain representation involves ensuring that the test accurately reflects content specifications, cognitive processes, etc. Inspecting domain relevance involves assessing the relevance of test items to the intended domain, as well as the overall relevance of the test to the assessment goals. Lastly, assessing construction appropriateness encompasses the extent to which process decisions are well-reasoned and quality controls are implemented (Sireci & Benítez, 2023; Sireci & Faulkner-Bond, 2014).

Content-based validity evidence is foundational for assessment validity, as it establishes the appropriateness of the test content for the assessed construct. Other validity sources, like internal structure-based evidence, can be compromised if test content is inadequate. Poorly sampled items or irrelevant content can distort scores and interpretations (AERA et al., 2014). Similarly, if test content has different effects on particular subgroups (e.g., based on socioeconomic status, race/ethnicity, or geography), scores may be biased (AERA et al., 2014; Gómez-Benito et al., 2018). In this article, we specifically propose a way of improving the assessment of item relevance using SMEs, although the framework proposed here may be generalizable to other types of SME-based evidence around test content (e.g., social desirability, wording).

The assessment of domain representation and relevance is often conducted either through a matching task or a rating task (Sireci & Faulkner-Bond, 2014). The former involves SMEs matching items to their closest domain and/or cognitive specifications (e.g., using Bloom's taxonomy). The latter comprises SMEs using ordinal scales to rate (a) the item representativeness/relevance for their respective domains, or (b) the similarity between items to assess whether the items correctly follow the expected domain structure (e.g., Li & Sireci, 2013).

Once the tasks are completed, the responses must be summarized using validity indices. These indices provide information to facilitate test developers in their decisions about the inclusion/exclusion of certain items. For instance, a common index for matching tasks is the proportion of times an item is assigned to its correct domain (Sireci & Faulkner-Bond, 2014). Low values indicate that the item has an unclear association with the expected domain. For rating tasks, several other summarization procedures or indices exist, such as Rovinelli and Hambleton's (1977) item-objective congruence index or multidimension scaling of item similarities (Li & Sireci, 2013). In alignment studies, evaluations often cover various aspects of assessments, like the extent to which target content areas and cognitive levels are accounted for in a test (Martone & Sireci, 2009). To measure alignment, matrix comparison indices are often used, quantifying how well test elements (e.g., number of items categorized by depth of knowledge and assessment content) align with curriculum standards (Porter, 2002).

Among the most commonly used and straightforward indicators in content validation, two stand out: the content validity index (CVI; Martuza, 1977) and Aiken's V index (Aiken, 1980). The CVI can be calculated at both item (I-CVI) and scale (S-CVI) levels. The I-CVI reflects the proportion of experts who agree on the relevance/representativeness of each item, being computed as the proportion of experts that consider an item relevant/representative (i.e., endorsing the higher half of the rating scale). The S-CVI is defined as the proportion of items about which all judges completely agree upon their relevance. In turn, Aiken's V consists of rescaling mean item evaluations to facilitate its interpretation:

$$V = \frac{\bar{x} - l}{K - l} \quad (1)$$

where $\bar{x}$ denotes the average rating for a given item, l is the lowest possible rating and K the highest rating option, thus providing a V value between 0 and 1. As it can be inferred from Equation 1, the I-CVI is a special case of Aiken's V for dichotomized data (e.g., relevant vs. irrelevant).

Regardless of the type of task, gathering SME judgment-based is a delicate matter. Tasks often involve a small group of experts (often from 3 to 20 SMEs; Almanasreh et al., 2019; McCoach et al., 2013; Rubio et al., 2003), so each judgment has great implications. Specifically, the adequacy of the assessments largely relies upon three key aspects: (1) expert selection, (2) task comprehension and engagement, and (3) result summarization indices. First, aiming to ensure a proper sample of experts, SMEs are often selected based on expertise criteria (e.g., years of experience). Second, clear instructions must be given, including detailed definitions of the domains in the test blueprint. Lastly, the choice for the appropriate validity index should provide accurate scores, leading to proper decisions about the items.

In realistic scenarios, where SMEs may have different backgrounds, their expertise, engagement with the task, or their use of the response scales (e.g., leniency/severity), may vary even among highly qualified experts (Sireci, 1998a). For instance, a more experienced rater may have a higher ability to discriminate

between lowly and highly relevant items. Severe experts can show a higher tendency to use the lower end of the response scale than lenient experts. For instance, a score of 4 for an SME whose average rating is 2 may not be comparable to a score of 4 for an expert with average ratings of 3.

In applied settings, SME performance and the rating precision are often disregarded. Consequently, when summarizing the results with traditional indices (e.g., with CVI or Aiken's V), these sources of measurement error are unaccounted for. In essence, despite the substantial advancements in measurement theory and methods in educational and psychological assessments (e.g., factor analysis, item response theory,), little has been translated into the context of SME-based validity assessment. In this sense, using statistical models to accommodate the SME variability may increase accuracy and generalizability of content validity scores, while also providing evidence on the appropriateness of the task (Rios & Wells, 2014).

In this matter, the item response theory (IRT) framework is especially suitable for such purposes. In fact, previous studies have accounted for rater variability using IRT models in educational assessments involving multiple graders (e.g., Lunz et al., 1994; Robitzsch and Steinfeld, 2018; Wu, 2017). It has been found, for instance, that considering the variability of graders' severity can have an impact on measurement accuracy and affect decisions regarding pass-fail outcomes (e.g., Lunz et al., 1990). In these applications, treating SMEs as measurement items in IRT models enables to estimate the discrimination and threshold parameters for each rater (e.g., if they correctly differentiate between relevant and irrelevant items, as well as their leniency/severity). This information provides evidence on the appropriateness of the experts, making it possible to detect suboptimal performance (e.g., SMEs that don't fully understand the task). As a result, it allows for the estimation of the relevance/representativeness scores of each item while considering SMEs' variability. Moreover, IRT provides a wide variety of evidence on the validity of the rating task, such as score reliability, standard errors, and goodness-of-fit indices. Accordingly, the goal of this article is to propose and illustrate the use of IRT modeling when summarizing content-related validity evidence from SMEs.

Within the IRT framework, the graded response model (GRM; Samejima, 1968) is a traditional polytomous response model. It is commonly applied in attitude and personality assessments, where Likert-type responses are frequent (e.g., Collado et al., 2015; Kreitchmann et al., 2019). The GRM models the probability of responding to each category of an item given the latent trait being measured. Specifically, the probability of endorsing category k or higher is a function of structural (i.e., item related) and incidental parameters (i.e., latent trait θ), as in Equation 2:

$$P(x \geq k) = \frac{1}{1+\exp\left[-a(\theta - b_{k\text{-}1})\right]} \tag{2}$$

Specifically, it involves the estimation of K parameters per item, namely one slope parameter a, and $K - 1$ threshold parameters b, being K the number of response categories. Usually, slope (or discrimination) parameters (a) represent the ability of an item to discriminate between different levels of the latent trait. For readers familiarized with factor analysis, discrimination parameters are equivalent to factor loadings given Equation 3:

$$\lambda = \frac{1.702^{-1} \cdot a}{\sqrt{1 + (1.702^{-1} \cdot a)^2}} \tag{3}$$

Threshold parameters represent the boundaries on the latent trait continuum where individuals are more likely to endorse category k or greater against previous categories. Each category, except the last one, has an associated threshold parameter, where $b_1 < \cdots < b_{K\text{-}1}$.

Traditionally, the input data for estimating the GRM consists of a matrix of dimensions N respondents × J items. With SMEs ratings, however, the data matrix takes the form of I evaluated objects (e.g., items) × M raters. As a result, the GRM parameters assume distinct interpretations within this context. The slope parameter represents an SME's ability to differentiate between items that are relevant and irrelevant in measuring the intended construct. The latent parameter θ reflects the specific feature being evaluated by the SMEs (e.g., item relevance). Finally, the b parameters are associated with the SME's leniency or severity. They reflect the level of item representativeness or relevance demanded by an expert to assign category k or beyond with a probability of 0.50. These parameters offer insights into the individual judgments made by experts, indicating their willingness to endorse different response categories based on the item's perceived relevance. To illustrate these parameters, the results for two SMEs in an expert rating task with four response categories are represented in Figure 1.

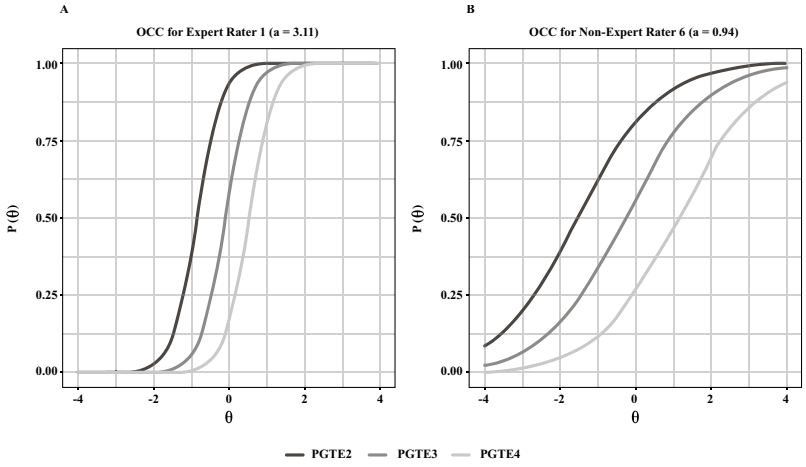
The rating task presented consisted in evaluating the relevance of 120 items for the measurement of the *Conscientiousness* domain. The response curves depicted in Figure 1, Panels A and B, represent item relevance ratings collected from two different raters: an expert rater with a Ph.D. in Psychology, and a non-expert rater with no degree in Psychology, respectively. As it can be observed, higher discrimination values correspond to steeper curves, indicating that smaller shifts in item relevance are required to move from category $k - 1$ to k. The threshold parameters were $b_{\text{expert}} = \{-0.83, -0.11, 0.54\}$ and $b_{\text{non-expert}} = \{-1.52, -0.27, 1.12\}$. These parameters are defined in θ (standardized normal) metric. The more centralized values of the expert rater indicate a tendency to endorse more extreme responses. For instance, in the case of b_1, items with a standardized relevance score lower than -0.83 are more likely to be rated as 1 out of 4 by the expert rater. Conversely, for the non-expert rater, a relevance score as low as -1.52 would already suggest a higher probability of assigning a rating greater than 1.

The probability of endorsing each specific category k can then be computed by calculating the differences between consecutive cumulative probabilities. As in Equation 4:

$$P(x=k) = P(x \leq k) - P(x \leq k\text{-}1) \tag{4}$$

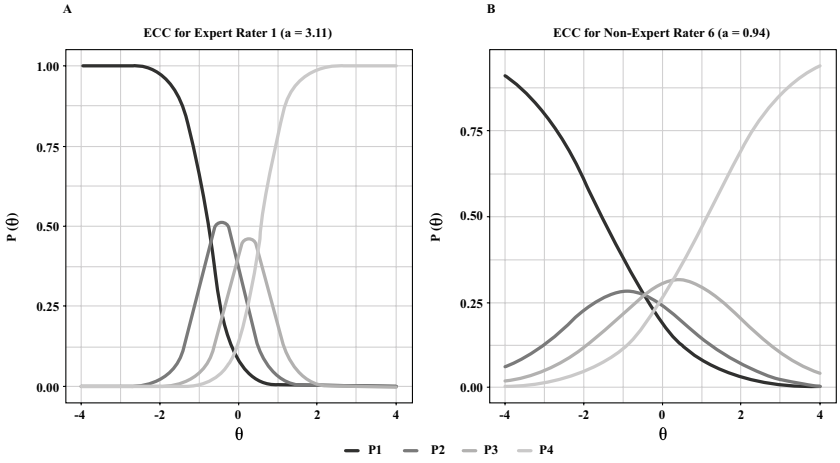
This transformation allows to visualize probabilities curves for the selection of each category independently. In the content validation context with SMEs, these curves can be referred to as *expert characteristic curves* (ECC), which reflect the relationship between item relevance and the rater's response probabilities. The corresponding ECCs of the previous example are shown in Figure 2.

Figure 1
Operating Characteristics Curves for Two Raters



Note. PGTEk: Probability that a category greater than or equal to k is chosen. Rater parameters are listed in Table 1.

Figure 2
Experts Characteristics Curves



Note. Rater parameters are listed in Table 1.

The scores derived from the GRM ($\hat{\theta}_{GRM}$) serve as an IRT-based content validity index representing SME-evaluated features such as item relevance. The $\hat{\theta}_{GRM}$ follows a standard normal distribution, allowing for comparisons between items and facilitating the selection of the most appropriate items. For absolute decisions, such as determining if an item is suitable or not (e.g., Aiken's V > 0.5), model-based expected responses can be computed using predicted probabilities for each judge (Equation 5). Subsequently, CVI and Aiken's V can be calculated using these expectations. With even-category ratings, a CVI or Aiken's V cutoff of 0.5 with expected responses parallels comparing $\hat{\theta}_{GRM}$ with the average central threshold across SMEs (e.g., $\bar{b}_2$ in four-category scales). I-CVI and Aiken's V based on model expectations (CVI_{GRM} and Aiken's V_{GRM}) will also be explored in this article.

$$E(x) = \sum_{k=1}^{K} k \cdot P(x = k) \quad (5)$$

To evaluate the utility of the GRM when gathering validity evidence based on the test content, a real-data study was conducted. This study focused on measuring the relevance of a set of personality items for assessing the *Conscientiousness* (CO) domain within the revised NEO Personality Inventory framework (NEO PI-R; Costa & McCrae, 1992). Three objectives are outlined as follows:

1. To evaluate the extent to which the parameter estimates obtained by using the graded response model can provide evidence regarding the adequacy of raters.
2. To investigate how the utilization of the IRT-based relevance scores improves the accuracy in determining whether an item measures or not a specific construct.
3. To assess the capability of IRT-based relevance scores in predicting the magnitude of factor loadings of the items.

For the second and third objectives, CVI and Aiken's V are used as a baseline for comparison. It is hypothesized that, since the IRT model allows for the consideration of each judge's discriminatory capacity and severity, the new proposal will lead to more precise item relevance indices, explaining a higher proportion of the variance of external criteria defined for objectives 2 and 3.

Method

The first part of the study involved administering a set of ninety-six CO items and twenty-four confounder items from other personality domains to expert and non-expert raters. Each group's ratings were analyzed using the graded response model. The average rater parameters were then compared between the two rater groups to provide evidence of the validity of the rater parameters (addressing Objective 1). In the subsequent phase, expert ratings were used to calculate the three content validity indices. These indices were compared in their ability to correctly distinguish between items designed to measure the CO domain and confounder items (Objective 2), and to predict the absolute standardized factor loadings of conscientiousness items in assessment data (Objective 3).

Participants

Item Relevance Ratings

The expert rater group comprised eight individuals with a Ph.D. degree in Psychology, with a mean age of 33.12 (*SD* = 4.82). Among the expert raters, 87.50% were male, while 12.5% were female. In contrast, the non-expert group consisted of a total of 13 individuals, of whom 53.85% were female. None of the members in the non-expert group possessed a university degree in Psychology. The average age of the non-expert group was 34.20 (*SD* = 14.27).

Personality Assessment Data

To estimate the standardized factor loadings (λ) of the items measuring conscientiousness, the dataset from Nieto et al. (2017) was also utilized. The original dataset included responses from 871 university students who responded to the 480-item pool, which also included twelve directed items aimed at detecting inattentive responding (e.g., *Please mark category five in this item*). All items were rated using a 5-point Likert scale, ranging from *strongly disagree to strongly agree*. The participants in the original dataset had a mean age of 19.99 (*SD* = 3.67). Among the participants, 19.61% were male, and 80.39% were female. These participants provided the basis for estimating the standardized factor loadings of the conscientiousness items in the current study.

Instruments

Personality Item Pool

A subset of a preexisting personality item bank originally developed and validated by Nieto et al. (2017) was used. The item bank consisted of 480 statements designed to measure the personality model within the NEO PI-R framework. Each item was specifically crafted to assess a general domain (neuroticism, extraversion, openness, agreeableness, and conscientiousness) and one of the six facets within each domain. For the current study, a total of 120 items were selected from the original item pool. This selection comprised 96 items that measured the Conscientiousness (CO) domain and an additional 24 items (6 items per domain) from the remaining domains. These additional items were included as confounders for the raters. Among the 96 conscientiousness items, fifteen items focused on dutifulness, seventeen items measured self-discipline, and sixteen items were associated with each of the other facets (achievement striving, competence, deliberation, and order). These item distributions were used to ensure comprehensive coverage of the conscientiousness domain and its specific facets within the study.

Procedure

The data collection procedures for the personality assessment data are detailed in Nieto et al. (2017). Participants were presented with two booklets containing items from the personality item pool, directed items, and the NEO-FFI-3. The administration of these booklets occurred over two one-hour sessions, with the order of presentation counterbalanced among participants. The data collection process took place within an official system at a Psychology faculty, where students received points for their participation.

For the rating task, raters were given a comprehensive definition of the *Conscientiousness* domain to ensure a clear understanding of the construct. This domain was described as encompassing features like dependability, organization, and goal-oriented behavior, among others, consistent with the NEO PI-R operationalization (Costa & McCrae, 1992). The rating task was carried out using an online survey platform. First, raters were shown examples of relevant and irrelevant items related to an unrelated trait (measuring sleeping difficulties). The participants were explicitly told to judge relevance without considering the item's key – both positively and negatively worded items could be considered relevant for measuring the domain. On each page, raters were reminded of the domain's definition and asked to rate the relevance of all 120 items based on this criterion using a 4-point rating scale, ranging from *not very relevant to very relevant.*

Data Analysis

Calculation of the Validity Indices

The calculation of the CVI and Aiken's V followed the procedure outlined in the Introduction section. For the CVI, items were considered relevant when they received a score of 3 or 4 from the experts, while non-relevant items were those with a score of 1 or 2. To estimate the scores for the graded response model ($\hat{\theta}_{GRM}$), a unidimensional model was fitted to the expert ratings using the *expectation-maximization* estimation method. The item relevance scores were then obtained through *expected-a-posteriori* estimation using the mirt package (Chalmers, 2012) in *R software enviroment* (R Core Team, 2023). Lastly, the model based CVI and Aiken's V (CVI_{GRM} and Aiken's V_{GRM}) were computed using the model expectations given $\hat{\theta}_{GRM}$ and the estimated rater parameters, as in Equation 5.

Personality Assessment Data Calibration

In Nieto's et al. (2017) original study, the items addressing each facet were calibrated separately under a unidimensional graded response model. For the purpose of the present study, the factor loadings associated with the CO domain were estimated using an exploratory bi-factor IRT model (e.g., Rios & Wells, 2014). This model included a general *conscientiousness* domain and its six facets (i.e., a total of seven factors were specified). The estimation was performed using the Metropolis-Hastings Robbins-Monro algorithm implemented in the *mirt* package and an orthogonal *bifactor* rotation (Jennrich & Bentler, 2011) was applied. This approach allowed for capturing the relevance of each item to the general construct of *conscientiousness*, rather than to its specific facets, which aligns with the instructions of the rating task. The calibrated model demonstrated a good fit to the data, with a RMSEA of 0.03 and a CFI of 0.96. All items had RMSEA values below 0.03. The empirical reliability of *conscientiousness* scores was 0.94, and ranged from 0.58 (*self-discipline*) to 0.90 (*deliberation*) for the facet scores. The average absolute standardized factor loading associated with the CO domain was 0.36 (*SD* = 0.17).

Comparison Criteria

The three validity indices were compared based on two criteria: the point-biserial correlations with a binary indicator on whether the items were designed to measure conscientiousness (MCO) or not, and, among the *conscientiousness* items, in their correlations with the absolute $\hat{\lambda}$ (i.e., $|\hat{\lambda}|$) from the personality assessment calibration. Additionally, the incremental validity offered by modeling expert ratings with the graded response model was assessed through the change in R^2 when including CVI_{GRM}, Aiken's V_{GRM}, or $\hat{\theta}_{GRM}$ in addition to either CVI or Aiken's V as predictors of $|\hat{\lambda}|$ and MCO in regression models. For predicting MCO, a logistic regression model was fitted and R^2 was calculated as the ratio of the predicted probabilities' variance to the outcome's variance (i.e., MCO). The significance of the incremental validity was assessed through the *likelihood ratio* test for the logistic models (i.e., with MCO as outcome), and the F-test for the linear models (i.e., with $|\hat{\lambda}|$ as outcome).

Results

Appropriateness of the Graded Response Model

The unidimensional graded response model had excellent fit with the expert ratings data (RMSEA=0.001,CFI=0.99) and slightly worse with the non-expert ratings data (RMSEA=0.063,CFI=0.95). Expert and non-expert parameters are presented in Table 1. As evidence on the validity of the rater parameters, average discrimination (i.e., *a*) parameters from the non-expert group were significantly lower than those from the expert group (t= -4.37,df=15.26,p<0.01). These differences suggest that expert raters are more capable of differentiating between items with low and high relevance in the CO domain than non-experts. Accordingly, expert ratings achieved a higher empirical reliability (0.92 for the eight experts) compared with 0.90 for the thirteen non-experts.

Also, average threshold (i.e., b_{avg}) parameters were lower for the expert group (t= -2.37,df=18.82,p=0.03), implying that, on average, expert raters endorsed higher grades for the items in this pool. This, coupled with the fact that 80% of the items were designed to measure *conscientiousness*, also provide evidence on the validity of the GRM model parameters.

Table 1
Rater Parameters for the Expert and Non-Expert Groups

	a	b_1	b_2	b_3	b_{avg}	λ
Expert group (Ph.D. in Psychology)						
Expert 1	3.11	-0.83	-0.11	0.54	-0.13	0.88
Expert 2	2.46	-0.28	0.24	0.59	0.18	0.82
Expert 3	2.97	-1.25	-0.22	0.81	-0.22	0.87
Expert 4	2.46	-0.92	-0.39	-0.21	-0.51	0.82
Expert 5	3.75	-0.92	-0.21	0.66	-0.16	0.91
Expert 6	2.10	-1.74	-0.67	0.10	-0.77	0.78
Expert 7	2.70	-1.36	-0.80	0.23	-0.64	0.85
Expert 8	3.69	-0.70	-0.02	0.51	-0.07	0.91
Mean	2.91	-1.00	-0.27	0.40	-0.29	0.85
Std. Dev.	0.59	0.45	0.34	0.34	0.32	0.05
Non-Expert Group (No Degree in Psychology)						
Non-expert 1	1.82	-0.90	-0.30	0.17	-0.34	0.73
Non-expert 2	1.63	-0.40	0.58	2.23	0.80	0.69
Non-expert 3	1.76	-0.29	0.11	0.68	0.16	0.72
Non-expert 4	2.58	-1.35	-0.76	-0.26	-0.79	0.83
Non-expert 5	1.78	-1.41	-0.70	0.57	-0.52	0.72
Non-expert 6	0.94	-1.52	-0.27	1.12	-0.22	0.48
Non-expert 7	0.77	-0.51	0.37	2.78	0.88	0.41
Non-expert 8	2.39	0.05	1.12	2.04	1.07	0.81
Non-expert 9	2.21	-0.90	0.10	1.20	0.13	0.79
Non-expert 10	1.95	-0.03	0.39	0.97	0.44	0.75
Non-expert 11	0.65	-1.16	0.08	2.27	0.40	0.35
Non-expert 12	2.19	-0.98	-0.26	0.18	-0.35	0.79
Non-expert 13	1.87	-0.37	0.51	2.07	0.74	0.74
Mean	1.73	-0.75	0.07	1.23	0.18	0.68
Std. Dev.	0.61	0.53	0.53	0.96	0.60	0.16

Note. b_{avg} = average of *b* parameters by rater; λ = discrimination transformed into factor loading metric.

Predictive Validity of Expert Ratings

All validity indices obtained high correlations with each other (Table 2). The correlation pattern reflects that Aiken's V is a middle-ground solution between the CVI and $\hat{\theta}_{GRM}$. However, despite the high similarity between the three measures, $\hat{\theta}_{GRM}$ obtained a higher correlation with the absolute standardized factor loadings ($|\hat{\lambda}|$; r=0.67), followed by Aiken's V_{GRM} (r=0.64). The CVI using the observed data obtained the lowest correlation (r=0.58). A similar pattern was found for the MCO, although with higher values and smaller differences between the measures ($0.62 \leq r \leq 0.67$). Notably, the computation of CVI and Aiken's V using model-based expectations performed better that the original indices using the observed responses.

In terms of predictive validity, the results showed that $\hat{\theta}_{GRM}$ performed the best, explaining 45% of the variance in $|\hat{\lambda}|$ and 57% of the variance in MCO (Table 3). The incremental validity of $\hat{\theta}_{GRM}$ over Aiken's V was also noteworthy, as it increased the proportion of explained variance by 11% for both $|\hat{\lambda}|$ and MCO, compared to the results obtained with Aiken's V alone. Consequently, when combining these two measures, they accounted for 49% and 62% of the variance in $|\hat{\lambda}|$ and MCO, respectively. The utilization of model-based expected responses in CVI and Aiken's V also demonstrated incremental validity, making it a highly favorable approach for situations that require making decisions about the relevance of the items in absolute terms.

Table 2
Correlations Between the Validity Measures and Criterion Variables

	CVI_{GRM}	Aiken's V	Aiken's V_{GRM}	$\hat{\theta}_{GRM}$	$\lvert\hat{\lambda}\rvert$	MCO
CVI	0.98	0.98	0.98	0.95	0.58	0.62
CVI_{GRM}		0.99	0.99	0.97	0.61	0.66
Aiken's V			0.99	0.98	0.61	0.65
Aiken's V_{GRM}				0.99	0.64	0.67
$\hat{\theta}_{GRM}$					0.67	0.67

Note. $\hat{\theta}_{GRM}$ = validity scores under the graded response model; $|\hat{\lambda}|$ = absolute standardized factor loadings of conscientiousness items; MCO = whether the item was designed to measure conscientiousness or not.

Table 3
Validity and Incremental Validity of the Indices

Predictor	$\lvert\hat{\lambda}\rvert$		MCO	
	R^2	ΔR^2	R^2	ΔR^2
CVI	0.34		0.44	
CVI + CVI_{GRM}	0.37	0.03*	0.57	0.13*
CVI + $\hat{\theta}_{GRM}$	0.46	0.12*	0.61	0.17*
Aiken's V	0.38		0.51	
Aiken's V + Aiken's V_{GRM}	0.45	0.08*	0.59	0.08*
Aiken's V + $\hat{\theta}_{GRM}$	0.49	0.11*	0.62	0.11*
CVI_{GRM}	0.37		0.52	
Aiken's V_{GRM}	0.41		0.55	
$\hat{\theta}_{GRM}$	0.45		0.57	

Note. CVI_{GRM} and Aiken's V_{GRM}= refers to these indices using model-based expectations under the graded response model. $\hat{\theta}_{GRM}$ = validity scores under the graded response model; $|\hat{\lambda}|$ = absolute standardized factor loadings; MCO = whether the item was designed to measure conscientiousness or not. *$p < 0.05$.

Discussion

This study provided new evidence on the utility of IRT modeling for SMEs ratings when gathering validity evidence based on test content. These findings align with prior research, which illustrates the importance of considering raters' variability when scoring essays (e.g., Lunz et al. 1994; Wu, 2017).

The main contributions of this study lie in the incremental validity of the IRT-based relevance scores compared to traditional CVI and Aiken's V. Firstly, the application of the graded response model has provided valuable insights into the appropriateness of the raters involved in the study. Specifically, the discrimination parameters of the raters differed significantly between the expert and non-expert groups. This finding shed light on potential issues in rater performance, such as lower levels of expertise or a lack of proper understanding of the instructions. By allowing to identify such problems, this study emphasizes the importance of considering rater contributions to the task at hand. Secondly, the IRT-based scores showed greater accuracy in determining item alignment with the specific construct, representing a significant improvement over traditional validity indices, especially when using the trait scores. By incorporating the IRT framework, this study offers a more precise approach to evaluating item relevance. The IRT relevance scores not only provided good predictions of the magnitude of the item loadings in a practical application but also outperformed the CVI and Aiken's V measures in terms of accuracy, providing also an improved model-based CVI and V for absolute decisions. It is worth noting that recent guidelines (e.g., Almanasreh et al., 2019) still consider these traditional indices as the preferred way to summarize this type of evidence, hence the importance of proposing an improved version of these indicators.

It is also important to consider some limitations of this study. Firstly, the estimation of GRM model parameters is linked to the item pool size, which may limit the generalizability of the findings to smaller item banks. In this sense, large item banks are frequently found in important fields of psychology, such as the assessment of normative or pathological personality (Abad et al., 2018; Oltmanns & Widiger, 2020). However, caution should be exercised when applying these results to contexts with limited item pools. In this regard, it is difficult to specify the exact conditions in which IRT could be an appropriate framework for rating tasks gathering content validity evidence. Among other factors, item pool size affects the precision of the rater parameters, which, in turn, may affect the estimation of the relevance scores (Thissen & Wainer, 1982). When assessing the adequacy of SMEs based on their discrimination parameters or characteristic curves, the standard errors or confidence intervals of raters' parameters should be considered. Nevertheless, even in situations where the item pool is limited and SMEs' parameters lack precision, a high number of raters has the potential to yield precise relevance scores. In essence, increasing the number of raters can compensate for poorly estimated SMEs' parameters, leading to accurate item scores. Secondly, related to the previous point, it is important to note that fewer or less discriminating SMEs may result in less reliable relevance scores. Nonetheless, one of the advantages of employing IRT with SMEs ratings is precisely the ability to estimate the reliability of the relevance scores. Thirdly, the IRT framework bases on a foundational assumption of conditional independence. To elaborate, it assumes that the responses of raters are independent from one another, and that the ratings provided by each SME for various items are also independent assessments. This underlying premise should be considered when applying IRT to SMEs ratings. For example, it is crucial that raters approach the rating task individually rather than collaboratively, and that the act of rating is oriented to a clear trait definition, rather than to item-to-item comparisons. Fourthly, in situations in which items have been well-developed and receive highly uniform ratings, a lack of variability can pose challenges in model estimation. When the variability of ratings is constrained, such as when all items are consistently rated as highly relevant, or when raters consistently exhibit extreme severity or leniency, the available information becomes limited. This limitation can give rise to calibration difficulties and ultimately result in reduced reliability of the relevance scores. Finally, the homogeneous high

discrimination parameters observed among all SMEs reflect their strong proficiency in effectively distinguishing between items. This outcome can be attributed to the stringent inclusion criterion, which mandated a PhD in psychology. It is important to acknowledge that if SMEs had more diverse backgrounds, a potentially more substantial effect could have been observed. Additionally, a wide variety of content validity or alignment indices exist (Martone & Sireci, 2009), and the suitability of IRT modeling should still be investigated for these other indicators.

In terms of future directions, the IRT-based approach holds potential for other applications beyond evaluating the relevance of items, such as assessing social desirability, or evaluating language appropriateness. By leveraging the advantages of implementing IRT modeling with SMEs ratings, several other benefits can be realized. One is the flexibility to use incomplete designs, where each SME rates only a subset of the items. This approach reduces the burden on SMEs by allowing them to focus on evaluating a more manageable number of items. In situations where there is a large pool of items to be assessed (e.g., Waugh et al., 2021), this feature becomes particularly valuable, as it optimizes the use of SME resources. Moreover, IRT modeling enables the implementation of adaptive assessments, wherein each SME rates only the items that align closely with their point of maximum discrimination. This adaptive approach tailors the rating process to the expertise and discriminatory abilities of each SME. For example, if an SME demonstrates exceptional proficiency in evaluating low relevance items, the adaptive assessment will prioritize presenting them with items from this range. This targeted administration of items could optimize the utilization of SMEs' expertise and enhances the precision and accuracy of the relevance assessments. The possibility of incorporating these features not only improves the efficiency and effectiveness of the rating process but also enhances the overall quality of the assessments. It allows for a more focused and tailored evaluation, ensuring that SMEs can provide their expertise in areas where they have the greatest impact. These approaches could potentially maximize the value of SMEs' contributions while minimizing their workload, making it a practical and efficient solution.

Author Contributions

Rodrigo Schames Kreitchmann: Conceptualization, Methodology, Formal Analysis, Writing - Original Draft, Writing – Reviewing and Editing. **Pablo Nájera:** Methodology, Formal Analysis, Writing - Original Draft. **Susana Sanz:** Methodology, Formal Analysis, Writing - Original Draft. **Miguel Ángel Sorrel:** Methodology, Formal Analysis, Writing - Original Draft.

Acknowledgements

We are thankful to Dr. Francisco J. Abad for his kind advices and helpful feedback during the development of this research.

Funding

This study has been partially funded by Universidad Autónoma de Madrid and Instituto de Ingenieria del Conocimiento, through the Chair in Psychometric Models and Applications. This funding source had no role in the design of this study, data collection, management, analysis, and interpretation of data, writing of the manuscript, and the decision to submit the manuscript for publication.

Declaration of Interests

The authors declare that there is no conflict of interest.

Data Availability Statement

R codes and data will be made available for reader upon request to the first author.

References

Abad, F. J., Sorrel, M. A., Garcia, L. F., & Aluja, A. (2018). Modeling general, specific, and method variance in personality measures: Results for ZKA-PQ and NEO-PI-R. *Assessment, 25*(8), 959–977. https://doi.org/10.1177/1073191116667547

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy, 15*(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] (Eds.). (2014). *Standards for educational and psychological testing* (14th ed.). American Educational Research Association.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with States' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21–29. https://doi.org/10.1111/j.1745-3992.2003.tb00134.x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Collado, S., Corraliza, J. A., & Sorrel, M. A. (2015). Spanish version of the Children's Ecological Behavior (CEB) scale. *Psicothema, 27*(1), 82–87. https://doi.org/10.7334/psicothema2014.117

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personaliry Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement, 7*(1), 3–13. https://doi.org/10.1177/014662168300700102

García, P. E., Díaz, J. O., & Torre, J. de la. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema, 26*(3), 372–377. https://doi.org/10.7334/psicothema2013.322

Gómez-Benito, J., Sireci, S., & Padilla, J.-L. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*, 104–109. https://doi.org/10.7334/psicothema2017.183

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76*(4), 537–549. https://doi.org/10.1007/s11336-011-9218-4

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of likert items. *Frontiers in Psychology, 10*, Article 2309. https://doi.org/10.3389/fpsyg.2019.02309

Li, X., & Sireci, S. G. (2013). A new method for analyzing content validity data using multidimensional scaling. *Educational and Psychological Measurement, 73*(3), 365–385. https://doi.org/10.1177/0013164412473825

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1994). Interjudge reliability and decision reproducibility. *Educational and Psychological Measurement, 54*(4), 913–925. https://doi.org/10.1177/0013164494054004007

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332–1361. https://doi.org/10.3102/0034654309341375

Martuza, V. R. (1977). *Applying norm-referenced and criterion-referenced measurement in education*. Allyn and Bacon.

Mastaglia, B., Toye, C., & Kristjanson, L. J. (2003). Ensuring content validity in instrument development: Challenges and innovative approaches. *Contemporary Nurse, 14*(3), 281–291. https://doi.org/10.5172/conu.14.3.281

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain: School and corporate applications*. Springer. https://doi.org/10.1007/978-1-4614-7135-6

Nájera, P., Abad, F. J., & Sorrel, M. A. (2021). Determining the number of attributes in cognitive diagnosis modeling. *Frontiers in Psychology, 12*, Article 614470.

Nieto, M. D., Abad, F. J., Hernández-Camacho, A., Garrido, L. E., Barrada, J. R., Aguado, D., & Olea, J. (2017). Calibrating a new item pool to adaptively assess the Big Five. *Psicothema, 29*(3), 390–395. https://doi.org/10.7334/psicothema2016.391

Oltmanns, J. R., & Widiger, T. A. (2020). The five-factor personality inventory for ICD-11: A aacet-level assessment of the ICD-11 trait model. *Psychological Assessment, 32*(1), 60–71. https://doi.org/10.1037/pas0000763

Penfield, R. D., & Giacobbi, Jr., Peter R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science, 8*(4), 213–225. https://doi.org/10.1207/s15327841mpee0804_3

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3–14. https://doi.org/10.3102/0013189X031007003

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108–116. https://doi.org/10.7334/psicothema2013.260

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling, 60*(1), 101–138.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research, 2*, 49–60.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94–104. https://doi.org/10.1093/swr/27.2.94

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100–100.

Sireci, S. G. (1998a). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299–321. https://doi.org/10.1207/s15326977ea0504_2

Sireci, S. G. (1998b). The construct of content validity. *Social Indicators Research, 45*(1/3), 83–117.

Sireci, S., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema, 35*(3), 217-226. https://doi.org/10.7334/psicothema2022.477

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100–107. https://doi.org/10.7334/psicothema2013.256

Tatsuoka, K. K. (1983). Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory. *Journal of Educational Measurement, 20*(4), 345–354.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*(4), 397-412. https://doi.org/10.1007/BF02293705

Waugh, M. H., McClain, C. M., Mariotti, E. C., Mulay, A. L., DeVore, E. N., Lenger, K. A., Russell, A. N., Florimbio, A. R., Lewis, K. C., Ridenour, J. M., & Beevers, L. G. (2021). Comparative content analysis of self-report scales for level of personality functioning. *Journal of Personality Assessment, 103*, 161–173. https://doi.org/10.1080/00223891.2019.1705464

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7–25. https://doi.org/10.1080/08957340709336728

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling, 59*(4), 453–470.