The value of using test response data for content validity: An application of the bifactor–MIRT to a nursing knowledge test

Article *in* Nurse Education Today · May 2015

DOI: 10.1016/j.nedt.2015.05.014

CITATIONS
5

READS
134

1 author:

Yuyang Cai
Shanghai University of International Business and Economics
43 PUBLICATIONS **325** CITATIONS

SEE PROFILE

Nurse Education Today xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Nurse Education Today

journal homepage: www.elsevier.com/nedt

The value of using test responses data for content validity: An application of the bifactor-MIRT to a nursing knowledge test

Yuyang Cai *

Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, CF703, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Accepted 16 May 2015
Available online xxxx

Keywords:
Bifactor multidimensional item response theory (bifactor-MIRT)
Content validity
Subject-matter experts (SME)
Test response data

SUMMARY

Aim: This paper aimed 1) to argue for the values of using test response data for content validation, and b) to demonstrate this practice using bifactor-multidimensional item response theory (bifactor-MIRT) for nurse education.
Method: The Nursing Knowledge Test (NKT) response data by 1491 nurse students from China were used for demonstration. Based on the content structure assumed by subject-matter experts (SME), a bifactor-MIRT model was constructed and tested. This involved five steps: dimensionality assessment, local dependence detection, model specification, calibrating and unit weighting.
Results: Dimensionality assessment results confirmed the content structure assumed by SME. Through local dependence detection and calibrating (i.e., item parameter check), items suspected of contaminating content were detected and those producing substantive harm were removed or constrained. Finally, content contributions by items to the overall scale and to their subscales were obtained through unit weighting.
Conclusion: Deficiencies residing in SME for content validation must raise attention. The study suggests the value of modeling test response data to compensate these deficiencies. The theoretical implication is discussed.

Introduction

Educational tests are commonly used to measure students' nurse knowledge (Redsell et al., 2003). To ensure measurement quality, it is fundamental to provide evidence of content validity (AERA, APA and NCME, 1999, 2014) that ensures "the content of the test is congruent with testing purposes" (Sireci and Faulkner-Bond, 2014, p. 101). Empirical endeavors to content validity usually rely on the judgment of subject-matter experts (SME) (Sireci and Faulkner-Bond, 2014), a practice also prevalent in nurse education (Beckstead, 2009). While this can help us understand certain facets of content validity, it risks unsatisfactory results with the test content due to complications resided in human judgment and other test procedures (Embretson, 1983; Sireci, 1998a). To overcome this limitation, empirical researchers turn to test response data (e.g., Colton, 1993; D'Agostino et al., 2011). This new orientation, however, has not been recognized in nurse education. The current study aimed a) to argue for the values of using test response data for content validity; b) to introduce bifactor-multidimensional item response theory (bifactor-MIRT) as an optional model for this practice, and c) to show how to apply this approach to evaluate the content validity of a nursing knowledge test.

Content Validity Evaluation

In a typical practice for content validity evaluation, a panel of SME are asked to link each test item with the test objectives, to assess the relevancy of the items to the content prescribed in the objectives, and finally, to judge if the items adequately represent the behaviors related to the intended content (Sireci and Faulkner-Bond, 2014; Waltz et al., 2010). This application, however, bears limitations. First, what it examines is human judgment per se rather than test content (Beckstead, 2009; Hogan, 2013). Assumption made in this way risks two types of confounding variances: uncertainty in the scale used for judgment data collection and uncertainty in human intuition. While many methods have been introduced to minimize the effect from the former (e.g., Lawshe, 1975; Newman et al., 2013; Penfield, 2003; Wilson et al., 2012), no progress has been made on the latter (Sireci, 1998b). The second limitation relates to information granurity. In reality, a test constructed under a single theory would consist of multiple content domains (Johnston et al., 2014). For test stakeholders such as teachers and students, a discrepancy between individual items as well as different domains in the reported score would have serious implications for diagnosing student performance (Leighton and Gierl, 2007). Making this differentiation, however, has proven to be difficulty for the SME (Murphy et al., 2013). In the paucity of studies that does deal with this difference (see Biddle, 2005; Haynes et al., 1995), SME are asked to rate directly the importance of different subcomponents. These ratings are then reflected in the test scores by balancing the number of items within each domain. The real contributions of different content

* Tel.: +852 3400 3824.
E-mail address: sailor_cai@hotmail.com.

http://dx.doi.org/10.1016/j.nedt.2015.05.014
0260-6917/

domains, however, are neither necessarily the same nor determined by the number of items (Rico et al., 2012). More objective approach is in need.

An idealistic solution would be to use test response data (Deville and Prometric, 1996). The value of test response data for content validity has been argued for decades ago. Lennon (1956) sees content validity as the interaction between test content and test responses. Ebel (1956) emphasizes that the only way to understand what content a test actually measures is to take the test by oneself. Guion (1977) asserts in his guidelines for content validity evaluation, "The response content must be reliably observed and evaluated" (p.7). This interactive view can find resonance among many other validity theorists (Embretson, 1983; Messick, 1989, 1995; Sireci and Geisinger, 1992). In short, whether test content is appropriate or not is one issue, whether it can actually activate behaviors related to the intended content is another. While SME judgment has been merited for understanding the first issue, test response data can be used to understand both, especially the latter.

Use of test response data can be found in a few educational studies. Colton (1993) used multivariate generalizability theory to evaluate the domain representation of test specifications of the ACT Mathematics Test (American College Testing, 1989). Deville and Prometric (1996) extended the multidimensional scaling method to model student's self-ratings of language competence. Ding and Hershberger (2002) applied structural equation modeling to examine the content meaning of each item and to testify whether the items measured the intended content domains at different levels. D'Agostino and his colleagues (2011) used confirmatory factor analysis with the 2004 Arizona state high school mathematics test. More recently, Schönbrodt and Gerstenberg's (2012) used exploratory factor analysis to examine the content clusters of motive inventories. Regretfully, no such exploration can be found in nurse educational research.

In nurse education, a test is usually designed to measure multiple domains of nursing knowledge. The test format is usually single multiple choice with four or more options and student responses are coded dichotomously. To provide granular information for content validity, an appropriate model is indispensable. The next section recommends bifactor-multidimensional item response theory (bifactor-MIRT) as an optimal method for our situation. We are aware that many other methods such as those applied in studies discussed above would suit our situation. However, a detailed discussion about the merits of those models falls out of the scope of this study.

Bifactor-MIRT

Recently, bifactor-MIRT has been valued as an idealistic method to evaluate test validity (Li and Rupp, 2011). This approach conceptualizes test multidimensionality as a set of uncorrelated factors: a general ability factor underlying all items and several domain-specific factors underlying different item subsets. Accordingly, the relationship between the probability of correct response to an item, given the general ability factor, its domain-specific factor and item characteristics, is formulated as:

$$P(y = 1|\theta_0,\ \theta_s) = c + \frac{1-c}{1 + \exp\{-[d + a_0\theta_0 + a_s\theta_s]\}},$$

where θ_0 is the general factor, θ_s is the domain-specific factor, c is the guessing parameter (lower asymptote), d is the item intercept, a_0 is the discrimination parameter on the general factor, and a_s is the discrimination parameter on its domain-specific factor. These item parameters can be estimated using computational methods such as Bock–Aitkin (Bock and Aitkin, 1981), Bifactor EM (Gibbons and Hedeker, 1992; Cai et al., 2011a, 2011b), Adaptive Quadrature (Schilling and Bock, 2005) and Metropolis–Hastings Robbins–Monro (Cai, 2010a, 2010b).

Bifactor-MIRT is deemed beneficial for our situation for several reasons. It can be used with dichotomous data in a confirmatory way (Reckase, 2009) to test the content structure assumed by the SME. Using this method, detecting contaminating content under the test becomes feasible. Moreover, it can be used to differentiate the relative content contributions of items to the overall scale or to their subscales. The former can be realized through item discrimination check after calibrating; the latter can be computed by extracting out the eigenvalues of the 2 (factors) by n (item discriminations) matrix for test items within the same subtest.

Evaluating the Content Validity of the NKT Using Test Response Data

Data Source

The current study used the Nursing Knowledge Test (NKT) response data. The NKT was an instrument used in a larger project that examined the relationship between nursing knowledge and nursing English reading ability. It was designed to measure knowledge in four subject areas: gynecology nursing, pediatrics nursing, basic nursing and medial nursing. Each subject comprised a subtest and tapped by six multiple choice questions. The test was constructed by two experienced healthcare teachers, who used to be professional nurses. Items came from the retired questions of the China Nurse Entry Test, a national licensing exam for Chinese nurses. Before test construction, they were informed of the purpose of study and particular content domains to cover. A sample question (in English) is:

Normally, an infant's anterior fontanel closes at:

A. 10 to 12 months B. two years old C. 18 to 20 months D. 12 to 18 months

Participants involved 1491 second-year nurse students (1465 females and 26 males) from eight medical institutions in China. They were all aged between 18 and 22 at the time of data collection. Before field entry, the author obtained ethical approval from his host university and had consent forms signed by the participating institution leaders and all participating students.

Procedures of Assessing the Content Validity of the NKT

The evaluation involved two phases: dimensionality specification and bifactor-MIRT modeling. Based on the SME, the test was specified to have five uncorrelated content dimensions: one general content domain (i.e., general nursing knowledge) representing content shared by all items and four particular content domains, one each representing the content exclusive to knowledge in gynecology nursing, pediatrics nursing, basic nursing, and medical nursing, respectively. The second phase comprised of five statistical steps: 1) assessing dimensionality; 2) detecting local dependence (LD); 3) model specifying, 4) calibrating and 5) unit weighting. Step 1 aimed to test the content structure assumed by the SME. Step 2 was to examine potential contaminating content at individual item or item cluster (two or more items) level; Step 3 was to determine the appropriate number of item parameters for best model estimation. Step 4 was to obtain item estimates to be used to identify the relative importance of individual items. These estimates were then used again in Step 5 to compute the relative importance of individual items to the overall scale and to their own subscales. Steps 1 to 4 were computed using the IRTPRO (Cai et al., 2011a) and Step 5 was computed by hand. The following section presents these results.

Results

Dimensionality Assessment

The results for the dimensionality assessment are presented in Table 1. The ΔG^2s due to successively adding four more domain-specific factors in the order of gynecology nursing, pediatrics nursing, basic nursing, and medical nursing to the general factor of nursing

Please cite this article as: Cai, Y., The value of using test responses data for content validity: An application of the bifactor-MIRT to a nursing knowledge test, Nurse Educ. Today (2015), http://dx.doi.org/10.1016/j.nedt.2015.05.014

Table 1
Bifactor for-M2PL DA results.

Factor fors[a]	$-2LL(G^2)$	*df*	ΔG^2	Δdf	*p*
g	41,805.06	48	–	–	–
g + f1	41,718.67	54	86.39	6	.000
g + f1 + f2	41,529.80	60	188.87	6	.000
g + f1 + f2 + f3	41,483.43	66	46.37	6	.000
g + f1 + f2 + f3 + f4	41,462.26	72	21.17	6	.000

[a] g = the general factor for nursing knowledge (feature shared by all NKT items); f1 = the domain factor for gynecology nursing knowledge (feature shared exclusively by items N1 to N6); f2 = the domain factor for pediatrics nursing knowledge (feature shared exclusively by items N7 to N12); f3 = the domain factor for basic nursing knowledge, (feature shared exclusively by items N13 to N18); f4 = the domain factor for medical nursing knowledge (feature shared exclusively by items N19 to N24); $-2LL$ (G^2) = −2 times loglikelihood; ΔG^2 = change of deviance; Δdf = change of degree of freedom; p = significance level.

knowledge (and associated changes in degrees of freedom) were 86.4(6), 188.9 (6), 46.4 (6), and 21.2(6), respectively. All four improvements were significant at the .00 level.

The bifactor-M2PL with one general factor and four domain-specific factors produced a $-2LL$ (G^2) of 41,462.26 with 72 parameters freely estimated. Compared with $-2LL = 41{,}805.06$ with the 48 parameters freely estimated for the one-factor 2PL-IRT, the decrease in deviance of 342.8 (df = 24) was significant at the .00 level and supported the significance of the four domain-specific factors. This suggests that the one- to four-dimensional hypothesis must be rejected. That is to say, the test was five-dimensional.

Local Dependence (LD) Detection

To detect LD violation, the bifactor-two-parameter logistic item response model (bifactor-M2PL) was applied with five intended factors: one general content factor (i.e., general nursing knowledge) and four domain-specific factors (i.e., gynecology nursing, pediatrics nursing, basic nursing and medical nursing). This involved two particular treatments: a) performing the bifactor-M2PL on the whole test to identify and screen out items with negative discrimination estimates; and b) performing the bifactor-M2PL on the modified test to examine item pairs showing LD statistic larger than 10.0 (Chen and Thissen, 1997).

The first trial of the bifactor-M2PL model produced three negative discrimination estimates on the general factor: −0.83 (N14), −0.19 (N15) and −0.24 (N18). These items were then dropped and the LD statistics based on the modified scale was evaluated. The results showed only one statistic larger than 10.0 (LD $X^2 = 12.7$, between N1 and N10). Its potential damage to model estimation was examined by checking the item discrimination estimates. The discrimination estimates of items N1 and N10 on the general factor were 0.68 and 0.41 and those on their domain-specific factors were 1.11 and 1.28, respectively. As all estimates fell within the reasonable range between 0.00 and 3.00, trivial harm of content contamination was suggested (Hambleton et al., 1991). In the end, only items N14, N15 and N18 were dropped as they were possibly measuring unintended domains.

Model Specification

Using the selected twenty one items and the five assumed content dimensions, two bifactor-MIRT models (i.e., the bifactor-M1PL and -M2PL models) were performed. The fit indices are presented in Table 2. The simple model produced a $-2LL$ (G^2) of 37,086.67 with 26 parameters freely estimated. Adding parameter *a* (the M2PL model) produced a reduced G2 of 388.20 (df = 37, $p < .00$), showing that the latter was performing better in interpreting the test response data. The less simple model was then considered more appropriate.

Table 2
Bifactor for-MIRT model specification results for the NKT.

Solution	$-2LL(G^2)$	*df*	ΔG^2	Δdf	*p*
Bifactor for-M1PL	37,086.67	26	–	–	–
Bifactor for-M2PL	36,698.47	63	388.20	37	.000

Note: $-2LL = -2$ times loglikelihood; ΔG^2 = change of deviance; Δdf = change of degree of freedom; p = significance level.

Calibrating

The bifactor-M2PL model accounting for one general content factor and four domain-specific factors was performed on the revised test (without items N14, N15 and N18 and constraining the discrimination estimates of N12 and N13 on their corresponding domain-specific factors to be zeroes). The discrimination estimates on the general and domain-specific factors, the threshold estimates, and the standardized errors for these statistics are shown in Table 3.

Baker (2001) recommends ranges of item discrimination values between 0.00 and 0.64 as low, between 0.65 and 1.35 as moderate, between 1.36 and 1.70 as high, and above 1.70 as perfect. For estimates on the general factor, four items produced high values above 1.35 (the highest was 2.20 by N23), eight items produced moderate values between 0.66 (by N1) and 1.27 (by N19), and only two items showed low values of 0.53 (by N16) and 0.63 (by N4). For the domain factor Gynecology Nursing, three of the six calibrated items showed low (the lowest was 0.38 by N4) to moderate discrimination (the highest was 0.91 by N5 and N6). For Pediatrics Nursing, three of the five unconstrained items showed high discrimination (the highest was 1.35 by N8) and two showed very low discriminations (0.12 by N11 and 0.16 by N9). For Basic Nursing, the two unconstrained items showed one low discrimination of 0.64 by N17 and one moderate discrimination of 0.75 by N16. For Medical Nursing, only one item showed a moderate discrimination of 1.26 (by N22) and all other five items showed low discriminations ranging from 0.04 by N20 to 0.54 (by N21). In all, no extreme values regarding the discrimination estimates appeared. It was then decided to stop model revision and use the structure for further analysis.

Table 3
Five-dimensional bifactor-M2PL calibrating results.

Domain	Item	a_{gi}	s.e.	a_{di}	s.e.	d_i	s.e.
Gynecology nursing	N1	0.66	0.08	0.63	0.16	−0.62	0.07
	N2	1.06	0.12	0.51	0.17	2.17	0.11
	N3	1.73	0.18	0.71	0.43	2.11	0.20
	N4	0.63	0.08	0.38	0.09	0.39	0.06
	N5	0.94	0.09	0.91	0.21	0.27	0.07
	N6	0.79	0.09	0.91	0.11	0.27	0.06
Pediatrics nursing	N7	0.73	0.12	1.31	0.28	1.03	0.12
	N8	1.03	0.12	1.35	0.18	0.68	0.08
	N9	0.87	0.09	0.16	0.09	0.45	0.06
	N10	0.96	0.10	1.17	0.13	−0.15	0.07
	N11	1.03	0.09	0.12	0.08	−0.41	0.06
	N12	1.47	0.13	0.00	0.00	1.12	0.08
Basic nursing	N13	1.13	0.10	0.00	–	0.54	0.06
	N16	0.53	0.08	0.75	0.21	0.78	0.08
	N17	1.10	0.10	0.64	0.16	0.34	0.07
Medical nursing	N19	1.27	0.13	0.31	0.16	1.98	0.11
	N20	0.89	0.08	0.04	0.09	0.07	0.06
	N21	1.48	0.13	0.54	0.13	1.01	0.08
	N22	1.05	0.10	1.26	–	0.58	0.07
	N23	2.20	0.24	0.27	0.23	2.03	0.15
	N24	1.08	0.10	0.28	0.17	−0.53	0.06

a_{gi} = discrimination on the general factor for; a_{di} = discrimination on the domain factor for; d_i = threshold.

Please cite this article as: Cai, Y., The value of using test responses data for content validity: An application of the bifactor-MIRT to a nursing knowledge test, Nurse Educ. Today (2015), http://dx.doi.org/10.1016/j.nedt.2015.05.014

Unit Weighting

The purpose of unit weighting was to understand the relative content contributions of test items to the general factor (the overall scale) and to their particular content domains (subscales). In doing so, four separate 2 (vectors of discrimination estimates on the general content factor and on the domain-specific factor, respectively) × 6 (items) matrixes were set up, each representing one of the four subscales. Two eigenvalues for each matrix were then extracted, one representing the weighting for the general factor and the other for the domain-specific factor (see Table 4 for detailed information).

For Gynecology Nursing, the original discrimination matrix produced by the items could be transformed to the following two-by-two matrix: $\begin{bmatrix} 6.41 & 3.75 \\ 3.75 & 2.73 \end{bmatrix}$. The two eigenvalues of this matrix were extracted out as 8.75 and 0.39, respectively. The larger value of 8.75 corresponded to the vector of $\begin{bmatrix} .85 \\ .53 \end{bmatrix}$. Therefore, the scalar on the top was the weight for the general content factor and the one on the bottom was for the domain-specific factor. Hence, the relative contributions of the Gynecology Nursing items to the general content factor and to its domain-specific factor were 0.85 (72% of the total variance explained) and .53 (28% of the total variance explained), respectively.

In the same way, the relative contributions of the Pediatrics Nursing items to the general content factor and to its domain-specific factor could be obtained as 0.77 (59% of the total variance explained) and .64 (28% of the total variance explained), respectively. The two values for the Basic Nursing subtest were 0.92 (85% of the total variance explained) and 0.38 (15% of the total variance explained). Those for the Medical Nursing subtest were 0.97 (94% of the total variance explained) and 0.25 (6% of the total variance explained). Their contributions to the general factor can be ranked into, from large to small, Medical Nursing (94% of the total variance explained), Basic Nursing (92% total variance explained), Gynecology Nursing (85% of the total variance explained) and Pediatrics Nursing (59% of the total variance explained).

Discussion and Conclusion

The paper argued for two values of using test response data for content validity. On the one hand, it can produce objective evidence to triangulate assumptions by the SME; it can provide granular information beyond human judgment on the other. This information includes symbols of contaminating content and relative content contributions by test items to the overall scale and to their subscales. It then recommended bifactor-MIRT as an appropriate model and demonstrated its application with the Nursing Knowledge Test. First the test content structure assumed by the SME was recovered. This was followed by a series of analyses in the order of dimensionality assessment, local dependence detection, model specification, calibrating and unit weighting. Dimensionality assessment results were used to examine the validity of the SME assumption in general. Local dependence detection and calibrating were applied to locate places where contaminating content might hide. Unit weighting was to find content contributions of all items to the overall scale and to different subscales.

Table 4
Information for weighting the NKT subtests.

Domain	Item	a_g^2	a_i^2	$a_g{}^a a_i$	Matrix		Eigenvalues	Eigenvector
Gynecology nursing	N1	0.45	0.32	0.38				
	N2	1.14	0.14	0.41				
	N3	2.86	0.44	1.12	6.41	3.75	8.75	.85
	N4	0.40	0.12	0.21	3.75	2.73	0.39	.53
	N5	0.92	0.88	0.90				
	N6	0.64	0.83	0.73				
Pediatrics nursing	N7	0.55	1.77	0.98				
	N8	1.00	1.72	1.31				
	N9	0.72	0.03	0.15	6.33	376	8.88	.77
	N10	0.94	1.39	1.14	3.76	4.94	0.09	.64
	N11	1.04	0.03	0.16				
	N12	2.07	0.00	0.00				
Basic nursing	N13	1.23	0.00	0.00				
	N16	0.28	0.37	0.32	2.72	.99	3.13	.92
	N17	1.21	0.37	0.67	.99	.74	0.33	.38
Medical nursing	N19	1.61	0.05	0.28				
	N20	0.77	0.00	0.02				
	N21	2.31	0.35	0.90	11.79	2.81	12.5	0.97
	N22	0.98	0.90	0.94	2.81	1.46	0.75	0.25
	N23	4.93	0.01	0.27				
	N24	1.19	0.14	0.41				

[a] $\boldsymbol{a_g}$ = discrimination on the general factor for; $\boldsymbol{a_i}$ = discrimination on the domain factor for.

With respect to content structure, the dimensionality assessment results clearly showed that the measure was multidimensional and that the five uncorrelated content domains, namely, general nursing knowledge, gynecology nursing, pediatrics nursing, basic nursing and medical nursing knowledge, were responsible for content structure the test. The content structure assumed by the SME was hence confirmed. This showed that test response data, if modeled using a suitable psychometric model (e.g., the bifactor-MIRT), can benefit content validity evaluation by triangulating the evidence provided by the SME. Evidence of content validity is then anchored into the test responses and becomes more convincible (Embretson, 2007). Note this is just the general information we obtained from modeling test response data. The next section discusses granular information that was available from this approach.

Results of local dependence detection and item parameter estimations enabled us to locate places where contaminating content might hide. Local dependence refers to significant covariance between items after controlling the intended content domains (Chen and Thissen, 1997). Taking advantage of the bifactor-MIRT, this study was able to locate one pair of items showing local dependence: N1 (within Gynecology Nursing) and N10 (within Pediatrics Nursing). Substantively, this suggested that N1 and N10 were measuring factor(s) other than the general nursing knowledge combined with genecology or pediatrics nursing knowledge. However, to prevent from information loss, they were not deleted immediately. The potential harm was checked by examining the values of the item discrimination parameters (to be elaborated later). As a result, the harm of contaminating content was regarded as trivial and none of them were removed. After all, it would be too much to demand each test item only measure the constructs as pure as intended. In the case of the SME, it would be hard to make such kind of objective decision.

Results of item discrimination estimates enabled us to examine content noise at item level. Item discrimination by nature refers to the extent to which an item is relevant to its intended domain, a term analogous to the concept of correlation from classic test theory (Gorin and Embretson, 2008). A zero value (e.g., N12 on Pediatrics Nursing) indicates that the item is not measuring anything related to the intended content domain. A negative value (e.g., N18 on the general factor) means that there are contaminating factors underlying this item (Embretson and Reise, 2000). Drawing on this feature, one would be able to zoom into potential measurement noises to such a meticulous extent that would be difficult for SME to achieve. These noises, if leaving unattended, could lead to invalid test results interpretation (Embretson, 2007).

A more essential concern for content validity evaluation would relate to the relative importance of individual items to the overall scale (Leighton and Gierl, 2007) or to their particular domains (or subscales). While this may sound unfeasible for the SME (Sireci and Faulkner-Bond, 2014); it becomes true directly or indirectly through examining the item discrimination parameters available from modeling test response data. Using Baker's (2001) criteria, we found that except for three items (N14, N15 and N18), all others performed adequately in measuring their intended content factors. Among them, nineteen were found to represent the general factor moderately well (or even better). In

Please cite this article as: Cai, Y., The value of using test responses data for content validity: An application of the bifactor-MIRT to a nursing knowledge test, Nurse Educ. Today (2015), http://dx.doi.org/10.1016/j.nedt.2015.05.014

addition, after partitioning out the general factor variance, nineteen items measured their domain-specific content factors to non-trivial extents.

At the domain-specific level, a typical SME assumption would take the contribution of each content domain to the overall scale as equal (Bobko et al., 2007). The NKT was developed with exactly the same idea. All subtests were assumed to have equal weight and designed to have an equal number of six items. Neither of the assumed equity, however, was supported by the results of the study. These findings ring the bell to the prevalent practice that assumes equality to content contributions at subtest or individual levels. To ensure higher measurement quality, no effort must be spared to search for solutions such as modeling test response data for evidence.

Conclusion

The deficiencies of SME judgment for content validity should raise attention from two perspectives. It is unable to ensure that the content implemented in the test can actually activate test behaviors relevant to this content. Besides, it is difficult for the SME to judge the existence of contaminating content hidden under the test, or to determine the relative content contributions by individual items to the overall scale or to their subscales. These deficiencies, as shown in this study, can be compensated by modeling test response data using bifactor-MIRT. The plausibility of the content structure assumed by the SME can be evaluated through dimensionality assessment. Information regarding contaminating content is available from local dependence detection and item estimates check. The relative importance of individual items can be objectively determined using the unit weighting technique. The results of the study should be able to shed light on our understanding of content validity. Under content validation studies that only rely on human judgment for evidence is the acquiesced belief that rater consistence is content validity. The acquiescence has remained so long without being scrutinized. Future content validation studies need to pay attention to this issue and endeavor to explore the portion of intended content that can flow to other procedures of the test life.

Acknowledgments

This study was supported by Educational Testing Service (ETS) under the TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment, by Assessment Systems Corporation (ASC) under the Grants for Graduate Students in Psychological and Educational Measurement Programs, and by the Faculty of Education of the University of Hong Kong (HKU) under the Faculty Research Fund.

References

American College Testing, 1989. Preliminary Technical Manual for the Enhanced ACT Assessment. American College Testing, Iowa City, IA.
American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999. Standards for Educational and Psychological Testing. Americal Psychological Association, Washington, D.C.
American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014. Standards for Educational and Psychological Testing. Americal Psychological Association, Washington, D.C.
Baker, F.B., 2001. The Basics of Item Response Theory. 2nd edition. ERIC, USA.
Beckstead, J., 2009. Content validity is naughty. Int. J. Nurs. Stud. 46 (9), 1274–1283.
Biddle, D., 2005. Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing. 2nd ed. Gower Publishing, Aldershot, England.
Bobko, P., Roth, P.L., Buster, M.A., 2007. The usefulness of unit weights in creating composite scores a literature review, application to content validity, and meta-analysis. Organ. Res. Methods 10 (4), 689–709.
Bock, R.D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. Psychometrika 46, 443–459.
Cai, L., 2010a. High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. Psychometrika 75, 33–57.
Cai, L., 2010b. Metropolis–Hastings Robbins–monro algorithm for confirmatory item factor analysis. J. Educ. Behav. Stat. 35, 307–335.
Cai, L., du Toit, S.H.C., Thissen, D., 2011a. IRTPRO: Flexible, Multidimensional, Multiple Categorical IRT Modeling [Computer Software]. Scientific Software International, Chicago, IL.
Cai, L., Yang, J.S., Hansen, M., 2011b. Generalized full-information item bifactor analysis. Psychol. Methods 16 (3), 221–248.
Chen, W.H., Thissen, D., 1997. Local dependence indexes for item pairs using item response theory. J. Educ. Behav. Stat. 22 (3), 265–289.
Colton, D.A., 1993. A multivariate generalizability analysis of the 1989 and 1990 AAP mathematics test forms with respect to the table of specifications. American College Testing, Iowa City, IA.
D'Agostino, J., Karpinski, A., Welsh, M., 2011. A method to examine content domain structures. Int. J. Test. 11 (4), 295–307.
Deville, C.W., Prometric, S., 1996. An empirical link of content and construct validity evidence. Appl. Psychol. Meas. 20 (2), 127–139.
Ding, C.S., Hershberger, S.L., 2002. Assessing content validity and content equivalence using structural equation modeling. Struct. Equ. Model. 9 (2), 283–297.
Ebel, R.L., 1956. Obtaining and reporting evidence on content validity. Educ. Psychol. Meas. 16 (3), 269–282.
Embretson, S., 1983. Construct validity: construct representation versus nomothetic span. Psychol. Bull. 93 (1), 179–197.
Embretson, S., 2007. Construct validity: a universal validity system or just another test evaluation procedure? Educ. Res. 36 (8), 449–455.
Embretson, S., Reise, S.P., 2000. Item Response Theory for Psychologists. Lawrence Erlbaum, Mahwa, New Jersy, London.
Gibbons, R.D., Hedeker, D., 1992. Full-information item bi-factor analysis. Psychometrika 57, 423–436.
Gorin, J.S., Embretson, S., 2008. Item response theory and Rasch models. In: McKay, D. (Ed.), Handbook of Research Methods in Abnormal and Clinical Psychology. Sage, Newbury Park, CA, pp. 314–334.
Guion, R.M., 1977. Content validity—the source of my discontent. Appl. Psychol. Meas. 1 (1), 1–10.
Hambleton, R.K., Swaminathan, H., Rogers, J., 1991. Fundamentals of Item Response Theory. Sage, Newbury Park, CA.
Haynes, S., Richard, D., Kubany, E., 1995. Content validity in psychological assessment: a functional approach to concepts and methods. Psychol. Assess. 7 (3), 238–247.
Hogan, T., 2013. Psychological Testing: A Practical Introduction. Wiley, New Jersey.
Johnston, M., Dixon, D., Hart, J., Glidewell, L., Schröder, C., Pollard, B., 2014. Discriminant content validity: a quantitative methodology for assessing content of theory-based measures, with illustrative applications. Br. J. Health Psychol. 19 (2), 240–257.
Lawshe, C.H., 1975. A quantitative approach to content validity. Pers. Psychol. 28, 563–575.
Leighton, J.P., Gierl, M., 2007. Cognitive Diagnostic Assessment for Education: Theory and Applications. Cambridge University Press, Cambridge.
Lennon, R.T., 1956. Assumptions underlying the use of content validity. Educ. Psychol. Meas. 16, 294–304.
Li, Y., Rupp, A.A., 2011. Performance of the $S-\chi 2$ statistic for full-information bifactor models. Educ. Psychol. Meas. 71 (6), 986–1005.
Messick, S., 1989. Validity. In: Linn, R.L. (Ed.), Educational Measurement, 3rd ed. Macmillan, New York, pp. 13–103.
Messick, S., 1995. Validity of psychological assessment: validation of inferences from persons' response and performances as scientific inquiry into score meaning. Am. Psychol. 50 (9), 741–749.
Murphy, K.R., Deckert, P.J., Kinney, T.B., Kung, M.C., 2013. Subject matter expert judgments regarding the relative importance of competencies are not useful for choosing the test batteries that best predict performance. Int. J. Sel. Assess. 21 (4), 419–429.
Newman, I., Lim, J., Pineda, F., 2013. Content validity using a mixed methods approach its application and development through the use of a table of specifications methodology. J. Mixed Methods Res. 7 (3), 243–260.
Penfield, R.D., 2003. Application of the Breslow–Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. Alberta J. Educ. Res. 49, 231–243.
Reckase, M.D., 2009. Multidimensional Item Response Theory. Springer Verlag, London, New York.
Redsell, S.A., Hastings, A.M., Cheater, F.M., Fraser, R.C., 2003. Devising and establishing the face and content validity of explicit criteria of consultation competence in UK primary care nurses. Nurse Educ. Today 23 (4), 299–306.
Rico, E.D., Dios, H.C., Ruch, W., 2012. Content validity evidences in test development: an applied perspective. Int. J. Clin. Health Psychol. 12 (3), 449–460.
Schilling, S., Bock, R.D., 2005. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. Psychometrika 70 (3), 533–555.
Schönbrodt, F.D., Gerstenberg, F.X., 2012. An IRT analysis of motive questionnaires: the unified motive scales. J. Res. Pers. 46 (6), 725–742.
Sireci, S., 1998a. The construct of content validity. Soc. Indic. Res. 45 (1-3), 83–117.
Sireci, S., 1998b. Gathering and analyzing content validity data. Educ. Assess. 5 (4), 299–321.
Sireci, S., Faulkner-Bond, M., 2014. Validity evidence based on test content. Psicothema 26 (1), 100–107.
Sireci, S., Geisinger, K.F., 1992. Analyzing test content using cluster analysis and multidimensional scaling. Appl. Psychol. Meas. 16 (1), 17–31.
Waltz, C.F., Strickland, O., Lenz, E.R., 2010. Measurement in Nursing and Health Research. Springer, New York.
Wilson, F.R., Pan, W., Schumsky, D.A., 2012. Recalculation of the critical values for Lawshe's content validity ratio. Meas. Eval. Couns. Dev. 45 (3), 197–210.

Please cite this article as: Cai, Y., The value of using test responses data for content validity: An application of the bifactor-MIRT to a nursing knowledge test, Nurse Educ. Today (2015), http://dx.doi.org/10.1016/j.nedt.2015.05.014