Examining the inseparability of content knowledge from LSP reading ability: an approach combining bifactor-multidimensional item response theory and structural equation modeling

Article *in* Language Assessment Quarterly An International Journal · March 2018

DOI: 10.1080/15434303.2018.1451532

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

Antony Kunnan
59 PUBLICATIONS **1,697** CITATIONS

Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: http://www.tandfonline.com/loi/hlaq20

Examining the inseparability of content knowledge from LSP reading ability: an approach combining bifactor-multidimensional item response theory and structural equation modeling

Yuyang Cai & Antony John Kunnan

To cite this article: Yuyang Cai & Antony John Kunnan (2018): Examining the inseparability of content knowledge from LSP reading ability: an approach combining bifactor-multidimensional item response theory and structural equation modeling, Language Assessment Quarterly, DOI: 10.1080/15434303.2018.1451532

To link to this article: https://doi.org/10.1080/15434303.2018.1451532

Published online: 22 Mar 2018.

LANGUAGE ASSESSMENT QUARTERLY
https://doi.org/10.1080/15434303.2018.1451532

Examining the inseparability of content knowledge from LSP reading ability: an approach combining bifactor-multidimensional item response theory and structural equation modeling

Yuyang Cai[a] and Antony John Kunnan[b]

[a]Department of Curriculum and Instruction, The Education University of Hong Kong, New Territories, Hong Kong; [b]Department of English, University of Macau, Taipa, Macao

ABSTRACT
This study examined the separability of domain-general and domain-specific content knowledge from Language for Specific Purposes (LSP) reading ability. A pool of 1,491 nursing students in China participated by responding to a nursing English test and a nursing knowledge test. Primary data analysis involved four steps: (a) conducting a bifactor-multidimensional item response theory model (bifactor-MIRT) analysis to establish measurement validity for the assumed domain-general factor and domain-specific factors underlying each test and to compute bifactor-MIRT direct scores; (b) transforming the bifactor-MIRT scores into composite scores; (c) conducting a confirmatory factor analysis with the composite scores to reconstruct the orginal bifactor-MIRT models, and (d) conducting a structural equation modeling analysis to explore the relationship between nursing knowledge factors (domain-general and domain-specific) and the nursing English reading factors (domain-general and domain-specific). The results showed that the domain-specific passage factors were significantly correlated with their corresponding domain-specific nursing knowledge factors and that domain-general nursing knowledge significantly predicted the variance of the domain-general reading factor. Overall, we concluded that content knowledge is inseparable from LSP reading ability. The implications for understanding LSP ability and for LSP reading test scoring are discussed.

Introduction

Content knowledge, also known as specific purpose background knowledge (Douglas, 2000), refers to specialized language knowledge in a particular discipline (e.g., specialized medical and nursing knowledge for professional nurses), a construct bearing similar connotation to the "propositional or topical content" knowledge proposed by Purpura (2004, 2017) in testing language for general purposes. In the context of reading assessment, content knowledge plays an important role in determining language for specific purposes (LSP) reading performance, because it enables readers to contextualize information from the texts to be processed, and this contextualization eventually makes comprehension achievable (Kintsch, 2012; Kintsch & Van Dijk, 1978). Despite this recognized importance, content knowledge has mostly been conceptualized as a *contextual factor* in conventional language assessment models (Douglas, 2000, 2013; Fulcher, 2000). For instance, the dominant language assessment model of Communicative Language Ability (CLA; Bachman, 1990; Bachman & Palmer, 1996, 2010) explicitly recognizes language knowledge and strategic competence as its core constituents but regards content knowledge as a type of general background knowledge (or prior

CONTACT Yuyang Cai sailor_cai@hotmail.com Department of Curriculum and Instruction, The Education University of Hong Kong, New Territories, Hong Kong.

knowledge) that may or may not be involved when making inferences about language test performance.

This ambiguous view of the role of background knowledge in language ability has prevailed in the language testing literature for decades. Some researchers see measuring background knowledge for a language test as unnecessary because the CLA would allow language users to “manipulate language functions appropriately in a wide variety of ways” (Davies, 2001, p.143). This view, however, has been challenged by other scholars. This can be seen from the question Alderson (1981) asked: “How can one possibly avoid involving prior knowledge, since comprehension and presumably production must depend upon the prior existence of some set of knowledge?” (p. 127). Later, proposing a meaning-oriented conceptualization of language knowledge, Purpura (2004) emphasized that language ability should not be about only the linguistic resources of the language (i.e., grammatical forms), but also about the intended and implied meanings of communication. In fact, Purpura (2017) argued that excluding propositional or topical meaning from the language knowledge construct is like “having language ability with nothing to say” (p. 36).

Over the years the inseparability of content knowledge from language ability has been gradually recognized by language assessment researchers (Chapelle, Enright, & Jamieson, 2008; Douglas, 2013; Fulcher, 2000). In the field of LSP assessment, the serious concern with content knowledge is well reflected in Douglas’s (2000) book, in which he proposed the concept of LSP ability, and explicitly argued for the theoretical status of content knowledge in LSP assessment.

Abundant empirical findings have accumulated that can support the significant relationship between background knowledge and LSP performance (Clapham, 1996; Krekeler, 2006; Liu, Chiu, Lin, & Barrett, 2014; Kim & Elder, 2015; Huhta, Vogt, Johnson, Tulkki, & Hall, 2013). Besides, theoretical advances in other areas, such as applied linguistics (Krashen & Brown, 2007), cognitive linguistics (Faber, 2012), and educational psychology (Tapiero, 2007), have also pointed to the critical connection between content knowledge and language proficiency. Regardless of this strong trend, most of these advances have taken two directions. They are either based on argumentative statements or on observations of the general association between content knowledge and LSP language performance. More detailed and meticulous investigations have not been conducted to examine the inseparability of content knowledge from LSP reading ability by taking content knowledge as a composite concept that can be decomposed into different units (e.g., domain-general and domain-specific content knowledge) and to examine psychometrically the inseparability of these different units from LSP reading performance.

In the context of a reading test it is common practice to filter out variance due to test takers’ familiarity with certain topics (knowledge in a particular domain or domain-specific content knowledge). This is usually achieved by applying the so-called testlet response models (Wainer, Bradlow, & Wang, 2007), which consists of one general factor (representing construct-relevant reading ability) and *m*-testlet factors (where *m* represents the number of passages and a testlet factor represents domain-specific content knowledge). By only accounting for the general factor, it is believed that all construct-irrelevant content knowledge would be controlled, and the final test scores would be free of content-relevant information (Wainer *et al*., 2007). However, whether or to what extent these testlet factors represent domain-specific content knowledge (the issue of separability) or whether the general factor is free of content knowledge shared by all specific domains (the issue of inseparability) is generally not examined.

In the context of assessing ESP (or LSP with English as the target language), where all different passages come from the same general domain (e.g., clinical nursing), the issue of inseparability is more serious. Put another way, it is highly possible that in an ESP reading test containing *m* texts, the contents of the *m* texts may share some common features. If this is true, it would be difficult to apply a testlet response model to filter out all content knowledge, at least, the domain-general content knowlge shared by the *m* specific domains. However, few empirical studies have been conducted to examine the nature of the testlet factors and of the general reading factor in terms of content knowledge. This type of evidence would provide language testers essential information for validating claims about content knowledge in their reading test scores, whether they are for or against the idea of inseparability.

The goal of the current study, therefore, was to examine whether content knowledge can be empirically separated from LSP reading test performance. In particular, we examined the nature of the testlet factors and the nature of the general reading factor of a testlet respone model in terms of two types of content knowledge: domain-specific content knowledge and domain-general content knowledge. To this end, we set our study in a nursing English reading test, a test originally consisting of four texts, each addressing one topic in clinical nursing: gynaecological nursing, pediatric nuraing, basic nursing, and internal medicine nursing, respectively. Therefore, domain-specific content knowledge in our study refers to clinical nursing knowledge in gynecological nursing, pediatric nursing, basic nursing, and internal medicine nursing, respectively; domain-general content knowledge referred to clinical nursing knowledge common to the aforementioned four subjects.

We then related this nursing English measurement model to an external nursing knowledge measurement model constructed similarly with the 1-plus-4 factorial structure. Building on these two measurement models, we explored the relevance of passage factors to the domain-specific nursing knowledge factors and the relevance of the general reading factor to the domain-general nursing knowledge factor. The first exploration helped to understand the (in)separability of domain-specific content knowledge, and the second was to understand the (in)separability of domain-general content knowledge from LSP reading ability. Before embarking on this journey, we first visited current practice in using testlet response models to separate the content knowledge effect (presumably, domain-specific only) from the LSP reading score. After that, we provided a brief introduction to the bifactor-multidimensional item response theory models (bifactor-MIRT; a particular testlet response model) and then to the computation of bifactor-MIRT-based composite scores, the critical concept that we relied on to simplify the complex bifactor-MIRT models for structural equation modeling.

Review of The Literature

Testlet Use in Reading Assessment, Content Knowledge Effect. and its Separation

A common practice in reading assessment is to provide several passages and test the comprehension of each passage with a testlet, or “a group of items related to a single content area that is developed as a unit” (Wainer & Kiely, 1987, p. 190). While using testlets in reading assessment has advantages, such as saving test time and increasing authenticity (DeMars, 2012; Van Der Linden & Glas, 2000), it might also incur additional variance due to factors such as content knowledge specific to the passage (Chen & Thissen, 1997; Lee, 2004; Wainer & Kiely, 1987; Yen, 1993). Consequently, correct responses to a content-relevant testlet would depend not only on respondents’ reading proficiency, but also on their content knowledge about the passage topic (Paap & Veldkamp, 2012; Sireci, Thissen, & Wainer, 1991). This type of content-relevant information (or variance), according to most CLA view holders, is a source of noise that needs to be controlled when making inferences about language ability. In response to this need, different psychometric models have been developed to eliminate this type of presumably content-relevant but “construct-irrelevant” information. These models, known as testlet response models (e.g., Bradlow, Wainer, & Wang, 1999; Gibbons & Hedeker, 1992; Li, Bolt, & Fu, 2006), distinguish one general reading factor underlying all test items, from several domain-specific content knowledge factors (also, specific passage or testlet factors), each corresponding to one subscale (Steinberg & Thissen, 2013).

This *1* (general factor)-plus- *n* (specific passage factors) structure fits in well with the conventional perception of the psychometric dimensionality of a reading test using testlets. According to this idea, the general factor represents reading ability presumably free of content-relevant information, and the passage factors represent content-relevant information demanded by corresponding testlets (Wainer & Kiely, 1987). Assigning a reading score by accounting for only the general reading factor, one would be able to minimize the effect of content knowledge carried in the passage factors (DeMars, 2012). For decades this approach has been taken as the standard approach to separating content knowledge from reading scores reported by domestic and international language assessment agencies. Regardless of its popularity, the validity of using testlet response models to partition out the effect of content knowledge remains unchecked and still controversial. First, while testlet response models are effective in isolating testlet factors that are passage-specific, the nature of the testlet factors (in particular, its content-relevant feature) has rarely been assessed by using external variables measuring content knowledge. Second, because these testlet factors are passage-based, these factors should also include text-relevant information, such as linguistic (mostly lexical) features particular to the different texts (Paap & Veldkamp, 2012). More empirical evidence is needed to clarify the nature of the testlet factors to understand the actual efficiency of using testlet response models in separating this so-called content knowledge effect when making inferences about reading test scores.

The concern with the separability of content knowledge is more serious when dealing with LSP reading assessment. Unlike tests of reading for general purposes, an LSP reading assessment would require that all text content, though randomly sampled from a variety of subject domains, should fall into a common broad discipline. Text content sampled would most possibly share the same knowledge base, or domain-general content knowledge as we call it in this study. Because this unit of content knowledge manifests itself in a general factor, a testlet response model that is designed to detect particular passage factors would not be able to separate this content-relevant information from other elements embedded in the domain-general factor. Put another way, even if the portion of domain-specific content knowledge could be excluded from score reporting and hence from making inferences about LSP reading performance, it is not sure whether all content knowledge, for example, domain-general content knowledge, can also be separated from score assignment and interpreting.

Given the tricky nature of this content knowledge effect, it seems too simple and hasty for both sides of the debate to take content knowledge as a unified term and discuss its (in)separability without zooming into the concept of content knowledge. It seems appropriate then to assume the existence of two units of content knowledge affecting LSP reading testlet responses: the domain-specific content knowledge contained in the passage factors and domain-general content knowledge contained in the general reading factor. By relating the two types of reading factors (the general reading factor and passage factors) to an external criteria constructed in the similar way (i.e., several domain-specific content knowledge factors and one domain-general content knowledge factor), we would be able to explore whether the passage factors contain domain-specific content knowledge (i.e., whether part of the content knowledge is separable, e.g., pediatric nursing knowledge) and whether the general reading factor contains domain-general content knowledge (i.e., whether all content knowledge is separable, e.g., general nursing knowledge common to all of the four nursing subjects).

For this complex exploration it would be extremely challenging to conduct simultaneously two types of analyses in a single statistical model: one type pertains to multidimensional item response modeling with dichotomously recorded data to ensure measurement validity, and another relates to structural equation modeling mostly used with continuous data to explore relations between different variables. To solve this problem, we chose to compute composite scores at the measurement level before conducting structural analysis. The merit of conducting this two-layer analysis is that it allows optimal estimation by using computer programs particularly designed for each type of computation. The next part introduces the computation of composite scores based on the bifactor-multidimensional item response model (bifactor-MIRT), a particular formula of the testlet response models.

Bifactor-MIRT and Bifactor-MIRT-Based Composite Scores

The bifactor-MIRT model is rooted in the concept of unidimensional item response theory (UIRT). The UIRT model presumes a single latent trait or ability that predicts an individual's test performance, on the conditon of certain item characteristics (i.e., item difficulty, item discrimination and guessing) (van der Linen & Hambleton, 1997). A limitation with UIRT models is their inability to handle multidimensional empirical data. In response to this limitation, a series of MIRT models (e.g., McDonald, 2000; Reckase, 1985) has been developed by extending the UIRT models. These extended models posit that an individual's response to a test item is predicted by multiple thetas (θs) (abilities) and multiple item characteristics (i.e., multidimensional item difficulty, multidimensional item discrimination, and guessing). Compared with UIRT models, the MIRT models not only involve more parameters to estimate but also produce estimates that are less straightforward to understand. Unlike the UIRT direct scores that are immediately interpretable, the MIRT direct scores (or θs) are meaningless by themselves and should not be interpreted before they are transformed into composite scores, which is done by pooling the effects from the different dimensions (DeMars, 2013; Thissen, 2013). Similarly, transformations are also necessary to turn multiple item discrimination estimates and multiple item difficulty estimates into multidimensional item discrimination (MDISC) and difficulty (MDIFF) estimates before interepretation (see the Appendix for a more technical introduction).

One particular group of MIRT models, the testlet response models, are chracteristic of profiling the multually correlated multiple factors in conventional MIRT models into the 1 (domain-general factor) -plus-*n* (domain-specific factors) structures. In the transformed structure the 1 general factor is to capture the common feature shared by all MIRT factors, and the *n* factors aim to capture the features particular to each subdomain (e.g., Bradlow *et al.*, 1999; Gibbons & Hedeker, 1992; Li *et al.*, 2006; Wainer & Kiely, 1987). A prominent example of these transformed models is the bifactor-MIRT discovered by Gibbons and Hedeker (1992), later improved by Cai, Yang, and Hansen (2011) and Reise (2012). This bifactor MIRT model allows for free estimation of subscale weights to testlet items and non-correlation among the constructed factors, a feature that is unavailable from other testlet models.[1] These features make the bifactor-MIRT model particularly suitable for modeling the tests in our study, given its ability to map a clear distinction between the domain-general factor and the domain-specific factors.

Similar to other MIRT models, the direct scores produced by bifactor-MIRT are meaningless by themselves and need to be transformed for interpretation. Reckase (2009) proposes the reference composite score for each subscale that balances the general factor and the testlet factor (to be illustrated below in the Data Analysis). This study followed Reckase in deriving the bifactor-MIRT-based composite scores for both the NERT (Nursing English Reading Test) and the NKT (Nursing Knowledge Test). For the NERT, a composite score refers to the score obtained after appropriately weighting the contributions from the general reading factor and from the corresponding passage factor.

Under the paradigm of CLA the computation of composite scores would be unnecessary, because the purpose of constructing the passage factors was to absorb and eliminate this content-relevant information during score assignment. In our case that content knowledge is now assumed to be construct-relevant, computing composite scores to account for these testlet factors becomes indispensable.

These composite scores are important for our exploration in at least two ways. First, because these scores were continuous data derived from bifactor-MIRT modeling, they were more appropriate for regression analysis than the original dichotomously recorded data. Second, although each set of composite scores only had four data points, they carried the full information from the domain-general factor and from the domain-specific factors (as shown later by the loadings of the composite scores on the general factor and by the variances of the uniqueness terms during the CFA stage). It is this simplicity and information history that enabled us to compress the original five dimensional bifactor-MIRT models into their correponding single-factor measurement models. This

[1] A detailed discussion comparing these models is out of the scope of this study. Readers interested in this issue are directed to the work by Steinberg and Thissen (2013).

transformation dramatically reduced the computation load demanded on existing statistical softwares and hence made our exploration possible.

The Study

Research Questions

This study was guided by three research questions:

(1) *What is the dimensionality of the Nursing English Reading Test (NERT)? Or can the the Nursing English Reading Test be distinguished as a domain-general factor and four domain-specific factors (i.e., testlet or passage factors)?*
(2) *Are the reading testlet factors (represented by the uniqueness terms contained in the reading composite scores) relevant to the nursing knowledge testlet factors (represented by the uniqueness terms contained in the nursing knowledge composite scores)?*
(3) *Is the domain-general reading factor (represented by the factor underlying the reading composite scores) relevant to the domain-general nursing knowledge factor (represented by the factor underlying the nursing knowledge composite scores)?*

The first question was to validate the assumed factorial structure of the NERT. The second and third questions were to test the assumed (in)separability of domain-specific content knoweldge and domain-general content knowledge from LSP reading ability.

Participants

A pool of 1,491 second-year nursing students in three-year nursing programs volunteered to participate in the study (see Table 1 for details). They were from eight medical colleges in China: two from Northeast China (*n* = 372, 24.9% of the total sample), two from North China (*n* = 255, 17.1%), two from Center China (*n* = 234, 15.7%), and two from Southeast China (*n* = 630, 42.3%). All students were between 17 and 23 years old. In addition, an overwhelming majority of 1,453 (97.5% of the total size) students were female and only 38 (2.5%) were male. The overwhelming portion of female participants should not be a serious concern, given the female-dominated reality of the nursing profession in China. For their English language background, all participants had six years' experience studying English as a foreign language at middle school level, one year's experience studying general English at the postsecondary level, and a two-month

Table 1. Demographic Information of the Participants.

		NO.	Gender		Age			
Geographic Area	College	(%)	F	M	17–18	19–20	21–22	23>
Northeast China	1	221 (14.8%)	211 (14.2%)	10 (0.7%)	17 (1.1%)	111 (7.4%)	86 (5.8%)	7 (0.5%)
	2	151 (10.1%)	151 (10.1%)	0 (0.0%)	6 (0.4%)	57 (3.8%)	84 (5.6%)	4 (0.3%)
	Subtotal	372 (24.9%)	362 (24.3%)	10 (0.7%)	23 (1.5%)	168 (11.3%)	170 (11.4%)	11 (0.7%)
North China	3	169 (11.3%)	163 (10.9%)	6 (0.4%)	49 (3.3%)	72 (4.8%)	48 (3.2%)	0 (0.0%)
	4	86 (5.8%)	84 (5.6%)	2 (0.1%)	6 (0.4%)	59 (4.0%)	21 (1.4%)	0 (0.0%)
	Subtotal	255 (17.1%)	247 (16.6%)	8 (0.5%)	55 (3.7%)	131 (8.8%)	69 (4.6%)	0 (0.0%)
Central China	5	104 (7.0%)	96 (6.4%)	8 (0.5%)	59 (4.0%)	41 (2.7%)	4 (0.3%)	0 (0.0%)
	6	130 (8.7%)	129 (8.7%)	1 (0.1%)	27 (1.8%)	62 (4.2%)	39 (2.6%)	2 (0.1%)
	Subtotal	234 (15.7%)	225 (15.1%)	9 (0.6%)	86 (5.8%)	103 (6.9%)	43 (2.9%)	2 (0.1%)
Southeast & South China	7	556 (37.3%)	545 (36.6%)	11 (0.7%)	45 (3.0%)	444 (29.8%)	67 (4.5%)	0 (0.0%)
	8	74 (5.0%)	74 (5.0%)	0 (0.0%)	7 (0.5%)	67 (4.5%)	0 (0.0%)	0 (0.0%)
	Subtotal	630 (42.3%)	619 (41.5%)	11 (0.7%)	52 (3.5%)	511 (34.3%)	67 (4.5%)	0 (0.0%)
Total		1491 (100.0%)	1453 (97.5%)	38 (2.5%)	216 (14.5%)	913 (61.2%)	349 (23.4%)	13 (0.9%)

experience studying nursing English. Regardless, in China, these students are not generally considered as proficient English users by their English teachers.

Instruments

The Nursing English Reading Test (NERT)

The NERT was designed to measure nursing English reading ability. The test used retired items from the reading section of the Medical English Test System Level Two (METS-2), an intermediate level of the Medical English Test System (METS; METS, 2007). The METS is a project jointly sponsored by China National Educational Assessment Authority, China State Administration of Foreign Experts Affairs, China Medical Association, and China Nurse Association. It is a four-level licensing test battery (four separate tests) developed to measure nursing students' ability to use English in nursing practice. A METS reading section usually contains four passages covering content from different nursing subjects (e.g., gynecology, pediatrics, etc.) required by the national curriculum for nurse education (Ministry of Health, 2007). The METS is regarded as an LSP test toward the more specific end of Douglas (2000) specificity continuum because it was designed to measure language proficiency but also with attention to domain-general and domain-specific nursing knowledge (METS, 2007). The NERT used in our study contained four passages, each addressing one of the following nursing subjects: gynecological nursing, pediatric nursing, basic nursing, and internal medicine nursing. For each passage there were five multiple choice questions to test students' reading comprehension (see Table 2).

To help readers understand the language features of the four texts, we calculated the Coh-Metrix indices (Graesser, McNamara, Louwerse, & Cai, 2004) to capture linguistic features of the texts. Coh-Metrix is a computational software that provides researchers with over 200 indices reflecting text features ranging from superficial level features (e.g., average sentence length and Flesch Reading Ease readability) to deeper level features (e.g., cohesion relations, world knowledge, and language and discourse characteristics) (McNamara, Graesser, McCarthy, & Cai, 2014). Because of space restriction we listed only five Coh-Metrix indices that were spotted to be related to the variances of the four NERT passage factors: DRINE (verbs as infinitives or unmarked forms), PCTEMPp (Text Easability PC Temporality, percentile), PCTEMPz (Text Easability PC Temporality, z score), WRDFRQa (CELEX Log frequency for all words, mean), and WRDNOUN (count of nouns). Besides these, three traditional indices were included because of their popularity in LSP reading research (e.g., Alderson & Urquhart, 1985; Clapham, 1996): text length (DESWC), average number of words in a sentence (DESSL), and Flesch Reading Ease (RDFRE). These indices are shown in the lower part of Table 2.

The Nursing Knowledge Test (NKT)

The NKT (in Mandarin) contained four subtests, each including six items and measuring nursing knowledge in one of the four nursing subjects: gynecology nursing, pediatrics nursing, basic nursing, and internal medicine nursing. The test was developed with the assistance of two content teachers from the eight medical institues involved in this study. The NKT was delivered concurrently with the NERT to the same cohort of nursing students. Because detailed information about the NKT has been published elsewhere (see Y. Cai, 2015), this information is not repeated here.

Data Collection

Before data collection, ethics issues were reviewed by the first author's then-host university. Local school permits and participant agreement signatures were obtained from school administrators and students, respectively. Participants were told of the purpose, background, and general steps during

Table 2. NERT Texts, Specific Knowledge Domains, and Text Features.

	Passage 1	Passage 2	Passage 3	Passage 4
NERT Text				
Specific Domain	Gynecological Nursing	Pediatric Nursing	Basic Nursing	Internal Medicine Nursing
Question Item	RE1 RE2 RE3* RE4 RE5	RE6 RE7 RE8* RE9 RE10	RE11 RE12* RE13* RE14 RE15*	RE16* RE17 RE18 RE19 RE20*
Coh-Metric Indices				
DESWC	182	207	291	271
DESSL	12.13	13.80	15.32	15.06
RDFRE	54.17	70.67	66.84	48.24
DRINE	10.99	33.82	20.62	7.38
PCTEMPp	18.67	88.30	62.93	18.41
PCTEMPz	−0.89	1.19	0.34	−0.90
WRDFRQa	2.89	3.05	3.02	2.86
WRDNOUN	280.22	222.22	254.30	287.82

Notes.

1. The symbol * beside the NERT items indicates reading for implicit meanings; items without the star symbol indicate reading for explicit meanings.

2. Coh-Metrix indices and abbreviations:

DESWC = Number of words; DESSL = average sentence length; RDFRE = Flesch Reading Ease; DRINE = verbs as infinitives or unmarked forms; PCTEMPp = Text Easability PC Temporality (percentile); PCTEMPz = Text Easability PC Temporality (z score); WRDFRQa = CELEX Log frequency for all words (mean); WRDNOUN = count of nouns.

data collection before they were asked to take the NERT and the NKT. The total time allowed for the two tests was 60 minutes.

Data Analysis

Immediately after data entry, the data set (N = 1,598) was checked for missing values. Cases with missing responses in excess of 5% of the total were removed. This resulted in 1,491 cases for actual analysis. Furthermore, missing values in the data were treated as wrong and were assigned zero scores (Enders, 2010). Statistical analyses involved five phases. First, item-level analyses were conducted by using SPSS Version 20.0 for Windows (IBM Corporation, 1989–2011). This was to compute the means, standard deviations, skewness, kurtosis, and internal consistency for the NERT items. The purpose of item-level analyses was not to provide answers to any research questions but to assist in a preliminary description of the data set.

Second, bifactor-MIRT modeling was performed by using IRTPRO (Cai, du Toit, Thissen, 2011) on the NERT data to evaluate the appropriateness of using this model. This involved two treatments: dimensionality assessment and local dependence (LD) detection. Three particular steps were followed for the first treatment: (a) performing a unidimensional 2PL-IRT model[2] on the reading assessment and obtaining model fit indices, including the deviance and degrees of freedom (G^2/df), Akaike information criterion (AIC), and Bayesian information criterion (BIC); (b) adding four text-specific testlet factors (hereafter domain factors; i.e., Domains 1–4) successively and obtaining the same set of model fit indices; and (c) examining the model fit improvement of using the complex model. A better fit model is determined if a complex model incurs a significant value of G^2/df (p < .05) and a smaller value for both AIC and BIC (Cai, Thissen, & Du Toit, 2011).

To detect item pairs showing severe LD, the LD $\chi2$ statistics (Chen & Thissen, 1997) were consulted. These statistics are (approximately) standardized values computed by comparing the observed and expected frequencies of responses to individual items with those of all other items. An LD $\chi2$ statistic larger than 10.0 is considered to violate the LD assumption for the MIRT models (Cai *et al*., 2011).

Third, the weighting ratios of each subscale for the general factor and for the testlet factor were computed by using the approach recommended by Reckase (2009). This involved three steps: (a) obtaining 4 (the number of testlets) sets of 2 (1 for the general factor and one for the testlet factor)-by-j (j = the number of items within each testlet) discrimination matrices for each subscale; (b) transforming each of the 4 matrices to a 2-by-2 matrix; and (c) deriving the eigenvectors that could be used as the weighting ratios (composite loadings) for the general factor and for each testlet factor, respectively.

Fourth, confirmatory factor analysis (CFA) was performed on the four NERT composite scores to examine (a) whether the common variance shared by the four scores could fully reserve information from the general reading factor, (b) whether each of the four uniqueness terms could fully reserve informtion from their corresponding testlet factors; and (c) whether the relative importance of the five different components (one general factor and four uniqueness terms) mapped out by the CFA remained as the same as their corresponding components in the validated bifactor-MIRT model. This reading measurement model, together with the NKT measurement model constructed in the same way (see Y. Cai, 2015), works as the basis for the next step analysis.

Finally, a structural equation model was constructed by regressing the reading measurement model on the nursing knowledge measurement model to seek answers to our last two research questions. Where necessary, model modifications were made to make sure the overall model fit the data well. The potential covariances between uniqueness terms across the two measurement models

[2] To produce stable parameter estimates for a MIRT model, a sample size of 1,000 (Reckase, 2009) to 2,000 (Ackerman, 1994) is necessary. Because the study only had a sample size of 1,491, the present study took the restrictive criteria and constrained the guessing parameter to zero.

and the regression coefficient from the general nursing knowledge factor to the general reading factor served to answer our second and third research questions, respectively. The analyses for the last two steps were performed on Mplus Version 7.4 (Muthén and Muthén, 1998–2015) by using the robust weighted least squares (WLSMV) estimator. The overall fit was assessed by using the comparative fit index (CFI), Root Mean Square Error Approximation (RMSEA), and SRMR, as suggested by SEM scholars (Marsh & Hocevar, 1985; Jöreskog, Sörbom, Du Toit, 2001; Byrne, 2010). A model with a CFI value of 0.95 or above and a RMSEA or SRMR value of .05 or below was considered as good fit. The next section reports the results of the five-step analysis involving the NERT. Data analysis of the first three steps for the NKT was reported elsewhere (Y. Cai, 2015) and are not reported in this article.

Results

Descriptive Statistics, Distributions, and Reliability Estimates

Table 3 presents the descriptive and reliability statistics for the NERT items. Mean values ranged from 0.17 (RE11) to 0.53 (RE7), indicating a moderate range of item difficulty levels. The least challenging items were RE7 and RE18, with means of 0.53 and 0.52, respectively. Both items were intended to test the subskill of understanding explicit meanings of particular nursing procedures. The most difficult items were RE10, RE11, and RE12, with means of 0.25, 0.17, and 0.25, respectively. RE10 and RE11 were intended to measure the understanding of explicit meanings, while RE12 was to measure the understanding of implicit meanings. (See Purpura, 1998 for his discussion of reading to understand explicit/implicit meanings and 2004 for his account of endophoric/explicit meanings and exophoric/implicit meanings of text information.)

The standard deviations ranged from 0.38 (RE11) to 0.50 (RE4, RE5, RE7, RE17, RE18, and RE20). All NERT items had skewness and kurtosis values within the limits of ±2, suggesting reasonably normal distributions (Bachman, 2004; Bachman and Kunnan, 2005).

The overall reliability estimates of the NERT based on Cronbach's alpha was 0.73, indicating that a large portion of the variance was due to general and group factors (Cortina, 1993). This medium size reliability estimate indicated that other factors might also contribute to the variance of the whole

Table 3. Item Descriptive and Reliability Statistics for the NERT (*N* = 1,491).

Item	Mean	S.D.	Skewness	Kurtosis	Point-Biserial Correlation (item-based)	Item-Total Correlation (composite score based)
RE1	0.39	0.49	0.45	−1.80	0.27	
RE2	0.36	0.48	0.58	−1.67	0.36	
RE3	0.30	0.46	0.89	−1.20	0.28	.79
RE4	0.44	0.50	0.26	−1.94	0.46	
RE5	0.44	0.50	0.23	−1.95	0.15	(excluding RE5)
RE6	0.40	0.49	0.41	−1.83	0.26	
RE7	0.53	0.50	−0.11	−1.99	0.47	
RE8	0.41	0.49	0.36	−1.87	0.37	.93
RE9	0.40	0.49	0.43	−1.82	0.30	(excluding RE10)
RE10	0.25	0.43	1.14	−0.70	0.06	
RE11	0.17	0.38	1.74	1.02	−0.09	
RE12	0.25	0.43	1.18	−0.60	0.04	
RE13	0.41	0.49	0.36	−1.87	0.25	.82
RE14	0.35	0.48	0.64	−1.59	0.32	(excluding RE11 and RE12)
RE15	0.39	0.49	0.44	−1.81	0.37	
RE16	0.39	0.49	0.45	−1.80	0.42	
RE17	0.48	0.50	0.07	−2.00	0.48	
RE18	0.52	0.50	−0.07	−2.00	0.29	.78
RE19	0.41	0.49	0.38	−1.86	0.39	
RE20	0.46	0.50	0.14	−1.98	0.29	
Cronbach's alpha = .73						Composite score based alpha = .93

test (Cho & Kim, 2015). For the item-total point biserial correlation index, there were four values below the rule-of-thumb cutoff point of 0.20 ($r_{\text{pbi\_RE5}} = 0.15$, $r_{\text{pbi\_RE 10}} = 0.06$, $r_{\text{pbi\_RE11}} = -\ 0.09$, $r_{\text{pbi\_RE12}} = 0.04$), which indicated that these four items might not be able to discriminate students' NERT performance. Of them RE11 seemed to be even more problematic given its negative value. We then returned to the original test question and found the plausibility of distractors might have led to the negative value for RE11. Regardless of potential problems identified with these four items so far, no decision was made for item-deletion at this stage in case of unnecessary information loss. Nonetheless, these results were used as important reference for decision making during subsequent bifactor-MIRT analysis.

Results of Bifactor-MIRT

Dimensionality Assessment

The dimensionality assessment of the NERT involved three steps: (a) performing a one-dimensional M2PL model with the test and obtaining the deviance and degrees of freedom; (b) adding the four domain factors F1 to F4 (or domain factors) one at a time and obtaining the significance of model improvement achieved due to the added factor; and (c) testing the significance of chi-square decrease due to the added domain factor, with the additional number of parameters that were estimated because of applying a complex model as the degrees of freedom. Results of dimensionality assessment based on the MH-RM estimation method (Cai, 2010a, 2010b) are presented in Table 4.

Results of model comparisons showed that adding, one by one, the four domain-specific factors F1 (gynecological nursing), F2 (pediatric nursing), F3 (basic nursing knowledge), and F4 (internal medicine nursing) yielded reduced −2LL[3] values (with degrees of freedom) of 112.32 (*df*= 5), 64.99 (*df*= 5), 103.66 (*df*= 5), and 97.23 (*df*= 5), respectively, all significant at the .00 level (i.e., more complex models were significantly better than less complex models). Besides, all complex models showed smaller AICs and BICs corresponding to their simple models, ranging from 68.54 to 101.59 for AICs and from 42.00 to 75.06 for BICs, respectively (see Table 4). These results indicate that the one- to four-dimensional hypotheses must be rejected. Put another way, the five-factor model provided comparatively better fit.

Local Dependence (LD) Detection

To detect items potentially measuring unintended factors, a bifactor-M2PL structure with one general nursing English reading ability factor and four passage factors was performed by using the Bock–Aitkin approach with expectation–maximization algorithm (Bock & Aitkin, 1981) on the 20 items of the NERT. The discrimination estimates of the items on the general nursing English reading ability factor were examined first before examining estimates on the testlet factors. The first trial of bifactor-M2PL yielded only one negative discrimination estimate ($a_{pRE11}= -0.36$) on the general nursing English reading ability factor. Because a negative discrimination estimate is abnormal and

Table 4. Bifactor-M2PL DA Results for the NERT

Factor*	−2LL(G^2)	*df*	ΔG^2	Δ*df*	*p*	AIC	ΔAIC	BIC	ΔBIC
P	36450.66	40	–	–	–	36530.74	–	36743.03	–
P+F1	36338.34	45	112.32	5	.000	36429.15	101.59	36667.97	75.06
P+F1+F2	36273.35	50	64.99	5	.000	36360.61	68.54	36625.97	42.00
P+F1+F2+F3	36169.69	55	103.66	5	.000	36266.23	94.38	36558.18	67.79
P+F1+F2+F3+F4	36072.46	60	97.23	5	.000	36185.40	80.83	36503.83	54.35

**Note*. P = the general factor; F1: testlet effect factor for Text 1; F2: testlet effect factor for Text 2; F3: testlet effect factor for Text 3; F4: testlet effect factor for Text 4. −2LL (G^2)= −2 times loglikelihood; ΔG^2= change of deviance; Δ*df*= change of degree of freedom; *p* = significance level; AIC = Akaike information criterion; BIC = Bayesian information criterion.

[3] In statistics, −2LL indicates the difference of model-data fit between a more complex model and a simpler model. A more complex model is preferred when the probability (or *p*-value) of this difference is significant.

usually indicates some unknown factors affecting response to the item, Item RE11 was dropped from further modeling. A second trial on the modified scale gave six other negative discrimination estimates on their respective domain factors: two on F1 (a_{1RE1}= −0.12, a_{1RE5}= −0.65), three on F2 (a_{2RE7}= −0.08, a_{2RE8} = −0.09, a_{2RE10} = −1.49), and one on F3 (a_{3RE13}= −1.40). Looking closer at these items, we found that RE1, RE5, RE7, and RE10 shared the same feature in using the TRUE or NOT TRUE statements in their question stems and suspected that there might be some question type effect. We also suspected that there might be some other idiosyncratic features with items RE8 and RE13. It seemed retaining these items in the model would provide no useful information relevant to domain-specific content knowledge. However, because these items did performance sufficiently in measuring the general nursing English reading factor, completely dropping them would lead to information loss. As a compromise, these items (except for RE7)[4] were retained in the model but with their discrimination estimates on their passage factors constrained. A third trial was then performed, and this produced no negative discrimination estimate. We then decided to move to the LD detection procedure.

The calibrated output after the third trial provided eight LD statistics larger than 10.0. From the largest to the smallest, they were 32.5 (related to RE5 and RE20), 17.5 (related to RE6 and RE17), 16.4 (related to RE13 and RE19), 13.2 (related to RE18 and RE19), 12.4 (related to RE8 and RE20), 12.0 (related to RE5 and RE17), and 10.1 (related to RE12 and RE17). The impact of LD on discrimination estimation was then examined. According to the IRTPRO output, the discrimination estimates related to these items ranged from 0.14 (by RE12 on the general nursing English reading ability factor) to 1.74 (by RE20 on Testlet Factor 4). Given these reasonable values, there was not enough evidence to reject the hypothesis that the discrimination estimates were not inflated because of violations of the LD assumption. Combined with previous dimensionality assessment results, it was concluded that the NERT reflected a domain-general dimension of nursing English reading ability and four domain-specific passage factors that might account for domain-specific content knowledge (i.e., gynecology nursing, pediatrics nursing, emergency nursing, and internal medicine nursing) and other factors such as text features.

Calibration. The calibration started concurrently with the dimensionality assessment and LD detection treatments. In doing so, a bifactor-M2PL model accounting for one general nursing English reading factor and four passage factors was applied by using the MH-RM estimation method on the revised NERT (excluding RE5, RE10, RE11, and RE12 and constraining the discrimination of RE1, RE8, and RE13 on their passage factors to zero values). The discrimination, threshold estimates, and standardized errors for these statistics, the derived multidimensional discriminations (MDISC$_{\text{i}}$s or A$_{\text{i}}$s), and multidimensional difficulties (MDIFF$_{\text{i}}$s or B$_{\text{i}}$s) are shown in Table 5.

In terms of their ability to differentiate the general reading factor, three items were highly discriminating (the a_{i}s ranging from 1.40 by RE17 to 1.79 by RE7), and all remaining calibrated items were moderately discriminating (the a_{i}s ranging from 0.64 by RE6 to 1.26 by RE16). For Testlet 1, all unconstrained items produced moderate discrimination estimates (with a_{i}s values of 0.73, 1.31, and 0.86 by RE2, RE3, and RE4, respectively). For Testlet 2, all items estimates were low (with a_{i}s ranging from 0.26 by RE6 to 0.54 by RE9). Of the two unconstrained items within Testlet 3, RE14 produced a low discrimination of 0.56 and RE15 produced a moderate estimate of 0.69. For items within Testlet 4, one item showed high discriminating power: RE20 (a_{i}, = 1.51),[5] two items showed moderate discrimination: RE17 ($a_{\text{i}} = 1.00$)and RE18 ($a_i = 0.83$) , and the other two showed low discrimination: RE16 ($a_i = 0.17$)and RE19 ($a_{\text{i}} = 0.45$).

[4] To avoid information loss, we did not constrain these items simultaneously but by following two steps: First, we only constrained items with relatively large negative loadings (RE5, RE10, and RE13) and found RE8 switched to positive. We then constrained RE1 and RE8 and found RE8 still remained positive. This is possible, given the negative values might result from the test method effect and that, after other items were constrained, the confounding factor was removed, even RE was not constrained.

[5] We checked the item and found the question asking about the symptom of Migraine, which was not explicitly provided in the text. We concluded that it was mostly due to high demanding on content knowledge that has led to this high discrimination value on the domain factor.

Table 5. Five-Dimensional Bifactor-M2PL Calibrating Results for the NERT

Domain	Item	a_p	s.e.	a_{i}	s.e.	d_{i}	s.e.	A_{i}	B_{i}
Text 1	RE1	0.73	0.08	0.00	0.00	−0.49	0.06	0.73	0.67
	RE2	1.07	0.09	0.73	0.16	−0.73	0.07	1.30	0.56
	RE3	0.92	0.14	1.31	0.67	−1.24	0.27	1.60	0.77
	RE4	1.58	0.14	0.82	0.26	−0.33	0.07	1.78	0.19
Text 2	RE6	0.64	0.08	0.26	0.18	−0.45	0.06	0.69	0.65
	RE7	1.79	0.17	0.35	0.17	0.25	0.07	1.82	−0.14
	RE8	1.00	0.09	0.00	0.00	−0.41	0.06	1.00	0.41
	RE9	0.87	0.08	0.53	0.00	−0.51	0.06	1.02	0.50
Text 3	RE13	0.66	0.07	0.00	0.00	−0.38	0.06	0.66	0.58
	RE14	0.95	0.00	0.56	0.00	−0.77	0.00	1.10	0.70
	RE15	1.15	0.09	0.69	0.14	−0.57	0.07	1.34	0.43
Text 4	RE16	1.26	0.10	0.17	0.11	−0.55	0.06	1.27	0.43
	RE17	1.40	0.12	1.00	0.14	−0.05	0.07	1.72	0.03
	RE18	0.70	0.08	0.83	0.11	0.11	0.06	1.09	−0.10
	RE19	1.03	0.09	0.45	0.09	−0.45	0.06	1.12	0.40
	RE20	0.68	0.09	1.51	0.19	−0.21	0.07	1.66	0.13

**Note.* a_p= discrimination on the general factor; a_i= discrimination on the domain factor; d_i= threshold; A_{i}= MDISC$_{\text{i}}$= multidimensional discrimination estimate; B_i=MDIFF$_{\text{i}}$= multidimensional difficulty.

For the multidimensional discrimination estimates, 5 of 16 items produced high MDISCs larger than 1.35. Among them, RE7 gave the largest value of 1.82 and RE3 gave the lowest value of 1.60. All other items were moderately discriminating (with a_{i}s ranging from 0.66 by RE13 to 1.34 by RE15). These discrimination estimates results provided information regarding the strength of each calibrated item in measuring their intended factors. The MDIFF$_{\text{i}}$s in the last column shows that there were only two negative B_{i}s: RE7 ($B_{\text{i}} = -0.14$) and RE18 ($B_{\text{i}} = -0.10$). All other 14 items produced positive B_{i}s values that ranged from .03 (by RE17) to .77 (by RE3), showing that the NERT as a whole was relatively challenging for the students. Considering that all item estimates were within the acceptable range, the conclusion that the NERT items were performing well in measuring the composite trait of nursing English reading ability could not be rejected.

Table 6. Information for Computing the NERT Composites.

Domain	Item	a_p^2	a_i^2	$a_p * a_i$	Matrix	Eigenvalues	Eigenvector
Text 1	RE1	0.53	0.00	0.00	5.02 3.28	7.41	.81
	RE2	1.14	0.53	0.78	3.28 2.92	0.52	.59
	RE3	0.85	1.72	1.21			
	RE4	2.50	0.67	1.30			
Text 2	RE6	0.41	0.07	0.17	5.37 1.25	5.67	0.97
	RE7	3.20	0.12	0.63	1.25 0.47	0.17	0.23
	RE8	1.00	0.00	0.00			
	RE9	0.76	0.28	0.46			
Text 3	RE13	0.44	0.00	0.00	2.66 1.33	3.35	0.89
	RE14	0.90	0.31	0.53	1.33 0.79	0.10	0.46
	RE15	1.32	0.48	0.79			
Text 4	RE16	1.59	0.03	0.21	5.56 3.69	8.63	0.77
	RE17	1.96	1.00	1.40	3.69 4.20	1.13	0.64
	RE18	0.49	0.69	0.58			
	RE19	1.06	0.20	0.46			
	RE20	0.46	2.28	1.03			

**Note.* a_p= discrimination on the general factor; a_i= discrimination on the domain factor.

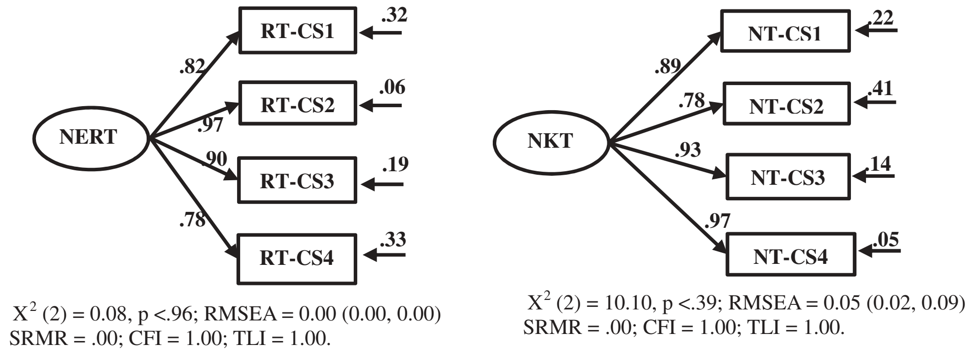
Results of Composite Scores Computation

This part reports the results of composite scores computation for the NERT (see Table 6 for detailed information). For Testlet 1, the original discrimination matrix produced by the four calibrated items was transformed into the following 2-by-2 matrix: $\begin{bmatrix} 5.02 & 3.28 \\ 3.28 & 2.92 \end{bmatrix}$. The two eigenvalues extracted out of this matrix were 7.41 and 0.52. The larger value of 7.41 corresponded to the vector of $\begin{bmatrix} .81 \\ .59 \end{bmatrix}$. Therefore, the scalar at the top was taken as the weight for the domain-general reading factor (gr), and the one at the bottom as the weight for the domain-specific passage factor (or Testlet 1 factor; t1). Hence, the composite score representing the contribution of the four calibrated Testlet 1 items (T1) can be obtained through the following equation: $\text{Composite}_{T1} = 0.81{*}\text{Score}_{gr} + 0.59{*}\text{Score}_{t1.}$

By using the same approach, the composite scores for the other three domains were obtained through the following equations: $\text{Composite}_{T2} = 0.97{*}\text{Score}_{gr} + 0.23{*}\text{Score}_{t2}$ for Testlet 2, $\text{Composite}_{T3} = 0.89{*}\text{Score}_{gr} + 0.46{*}\text{Score}_{t3}$ for Testlet 3, and $\text{Composite}_{T4} = 0.77{*}\text{Score}_{gr} + 0.64{*}\text{Score}_{t4}$ for Testlet 4.

Results of CFA

This part reports the results of the confirmatory factor analysis (CFA) with the NERT composite scores and with the NKT composite scores, respectively. A single round of confirmatory factor analysis with each data set showed that the simple measurement structure fit well with each set of composite scores (see Figure 1(a) for the NERT model and Figure 1(b) for the NKT measurement model). The NERT measurement model fit the data almost perfectly (X^2 (2) = 0.08, p < .96; *RMSEA* = 0.00 (0.00, 0.00); *SRMR* = 0.00; TLI = 1.00; CFI = 1.00). For standardized loadings on the general factor, the values corresponding to the four composite indicators were 0.82, 0.97, 0.90, and 0.78, remaining exactly the same as they were used for computing these composite scores. The same results emerged for the variances of the uniqueness terms for the composite indicators, the values being 0.32, 0.06, 0.19, and 0.33, respectively. The NKT measurement model also fit the data perfectly (X^2 (2) = 10.10, p < .39; *RMSEA* = 0.05 (0.02, 0.09); *SRMR* = 0.00; TLI = 1.00; CFI = 1.00). For standardized loadings on the general factor, the values corresponding to the four composite indicators were 0.89, 0.77, 0.93, and 0.97, remaining exactly the same as used for computing these composite scores if ignoring the rounding effects. Their corresponding uniqueness term variances



X^2 (2) = 0.08, p <.96; RMSEA = 0.00 (0.00, 0.00)
SRMR = .00; CFI = 1.00; TLI = 1.00.

Model 1.1: The NERT measurement model

X^2 (2) = 10.10, p <.39; RMSEA = 0.05 (0.02, 0.09)
SRMR = .00; CFI = 1.00; TLI = 1.00.

Model 1.2: The NKT measurement model

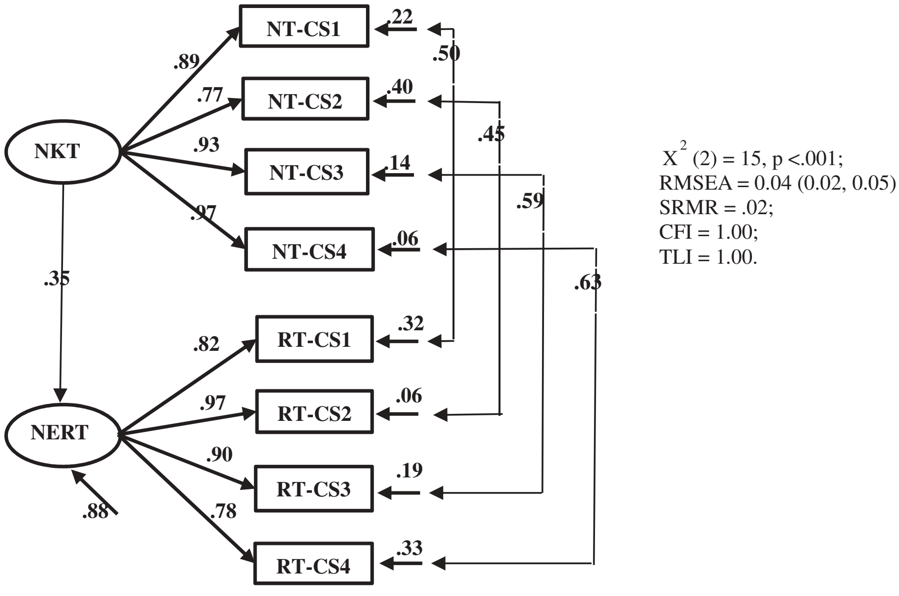
Figure 1. Individual CFA measurement models.
Acronyms. NERT = The nursing English reading test factor (representing domain-general LSP reading ability); RT-CS1 to RT-CS4 = composite scores for reading Text 1 to Text 4; NKT = The nursing knowledge test factor (representing domain-general/common clinical nursing knowledge); NT-CS1 to RT-CS4 = composite scores for nursing knowledge Subtest 1 to Subtest 4.

were 0.22, 0.41, 0.14, and 0.05, respectively. Building on these results, we confirmed that both the NERT and the NKT composite scores were valid representations of their corresponding bifactor-MIRT models. The next part reports results of performing a structural model using these two measurement models.

Results of SEM

To seek answers to our last two research questions, we regressed the NERT measurement model on the NKT measurement model with particular attention to potential associations across the two group of uniqueness terms. A first round model estimation produced indices indicating poor fit between the hypothesized model and the data (X^2 (19) = 1334.49, p < 0.00, *RMSEA* = 0.22 (0.21, 0.23), SRMR = 0.03, CFI = 0.89, TLI = 0.84). Large modifications indices suggested the need to release the zero parameter constraints on the four pairs of uniqueness terms. The largest value of 419.58 was related to the NERT Passage 3 and NKT Subtest 3, followed by 339.01 related to NERT Passage 1 and NKT Subtest 1, then by 323.51 related to NERT Passage 4 and NKT Subtest 4, and finally by 165.16 related to the NERT Passage 2 and NKT Subtest 2. Obviously, this automatic matching-up revealed that the modification suggestions were not simply an outcome of statistical tactics but a manifestation of substantive disposition embedded in the design of the two measurements.

These four pairs of covariance values (thereafter labeled as Pair 1 to Pair 4) were then added successively to perform another four rounds of model estimation (see Figure 2 for the diagram with the four modifications). After making the four modifications, the fit indices increased to excellent fit (x^2 (15) = 41.72, p < .001, RMSEA = 0.04 (0.02, 0.05), SRMR = 0.02, CFI = 1.00, TLI = 1.00). The values of the correlations for the four pairs of uniqueness terms were, from largest to the smallest, 0.63 by Pair 4 (addressing internal medicine nursing), 0.59 by Pair 3 (addressing emergency nursing), 0.50 by Pair 1 (addressing gynecology nursing), and 0.45 by Pair 2 (addressing pediatric



Model 2: The final structural model

Figure 2. The final structural model (Estimates Standardized).
Acronyms. NERT = The nursing English reading test factor (representing domain-general LSP reading ability); RT-CS1 to RT-CS4 = composite scores for reading Text 1 to Text 4; NKT = The nursing knowledge test factor (representing domain-general/common clinical nursing knowledge); NT-CS1 to RT-CS4 = composite scores for nursing knowledge Subtest 1 to Subtest 4.

nursing). All uniqueness covariances were significant at the level of p = .000. Drawing on these results, our hypothesis that the uniqueness variances of the reading composite scores are due to knowledge in specific nursing subjects cannot be falsified.

For the coefficient estimate of the path from the general nursing knowledge factor to the general nursing English reading factor, the value is 0.32 (R^2 = 10.2%). This indicated that, even after eliminating the passage-specific nursing knowledge carried in the testlets (i.e., knowledge exclusive to each of the four particular subjects of nursing knowledge: gynecology nursing, pediatrics nursing, emergency nursing, and internal medicine nursing), the general nursing English reading factor still contains a substantial portion of information from domain-general nursing knowledge shared by all four nursing subjects. That is to say, even if the domain-specific content knowledge effect could be successfully eliminated by removing those passage factors, there would still be a significant portion of variance due to domain-general content knowledge in the reported scores.

Discussion

According to *Standards of Psychological and Educational Measurement* (AERA, APA, & NCME, 2014), a fundamental element of test scoring is to ensure a substantive basis for selecting suitable scoring models and interpreting test scores. Practically, this is equal to presuming a construct structure (dimensionality) based on sound theory and testing it using competitive versions of a selected psychometric model. In our case the NERT was hypothesized to consist of five uncorrelated factors: a domain-general nursing English reading factor and four additional passage factors. The domain-general factor was assumed to represent the reading ability demanded by all reading test items, and each of the passage factors was assumed to mainly contain nursing knowledge exclusive to one of the particular nursing subjects: gynecology nursing, pediatrics nursing, basic nursing, and internal medicine nursing. This hypothesized structure was tested by using a series of bifactor-MIRT models. Results of bifactor-M2PL modeling supported our hypothesis that nursing students' performance on the NERT was determined by two types of factors: a domain-general nursing English reading factor and four passage factors.

The SEM results provided explanations for the content-relevant nature of the passage factors (addressed by the second research question) and of the general reading factor (addressed by the third research question). The automatic matching-up between the two groups of uniqueness terms (one group representing the passage factors and the other representing the nursing knowledge subtest factors) provided supporting evidence for the long-held assumption that content knowledge is a major factor in passage effect (Wainer *et al*., 2007; Wainer & Wang, 2000). The message conveyed in this finding is that, by deliberately ignoring the passage factors during score assignment, it is possible to exclude a portion of content-relevant information from score reporting. That is to say, the effect of content knowledge (domain-specific) is separable. This psychometric separability, however, should not directly lead to the overall conclusion of substantive separability. When making such a conceptual conclusion, one must also consider risks this practice might incur. Substantively, the salience of the four covariances in our study meant students with better nursing knowledge on a particular subject would have higher possibility of providing correct responses to the testlet corresponding to that subject. If this type of content knowledge effect were to be excluded for score assignment, we would have to adjust the higher scores due to better mastery of nursing knowledge to some lower values. Arguably, this practice is equivalent to punishing the cohort of higher scorers, who have "performed too well," in using the knowledge beneficial for their future career. Thus, in conceptualizing the role of content knowledge for LSP reading performance, we would need to account for the issue of test fairness (Kunnan, 2014, 2017). Otherwise, we would, sooner or later, hear laments similar to what was heard by Alderson and Urquhart's (1988): "The test is unfair, I am trained to be a nurse."

Another consequence of deliberately excluding testlet factors would relate to information loss. This concern can be partly confirmed by consulting the relationship between the variances of the uniqueness terms and the Coh-Metrix indices. As shown earlier, the variances of the four uniqueness terms for Text 1 to Text 4 were 0.32, 0.06, 0.19, and 0.33, respectively. Ranked by size from large to small, they followed the order of Text 4, Text 1, Text 3, and Text 2. An overview of the Coh-Metrix indices would show that this order is perfectly consistent with the ordering of WRDNOUN (the count of nouns) and with reverse orderings of other four indices: the DRINE (verbs as infinitives or unmarked forms), PCTEMPp (Text Easability PC Temporality, percentile), PCTEMPz (Text Easability PC Temporality, z score), and WRDFRQa (CELEX Log frequency for all words, mean). This consistency seemed to suggest that the testlet factors not only contained information of domain-specific content knowledge but also linguistic information particular to each passage. If this inference is true, deliberately excluding the passage factors for the sake of controlling content-relevant information might also result in the exclusion of the language knowledge effect. This would be similar to the notorious problem of "throwing out the baby with the bathwater."

The most important question raised is the relationship between the domain-general reading factor and the domain-general content knowledge factor. According to the SEM results, even by isolating the portion of domain-specific content knowledge effect into the passage factors, content knowledge still presented itself in the reading test scores in the form of domain-general content knowledge (r = .35, p < .001; R^2 = 12.25%). The finding provides direct evidence supporting the presumably psychometrical inseparability of content knowledge (i.e., domain-general) from LSP reading performance (Alderson, 1981; Douglas, 2000, 2013). And we would argue, more generally, from all reading performance. In line with Purpura (2017), the purpose of reading is to extract meaning, and meaning is composed of propositional information (content) that a reader builds a mental representation of.

Similar to the interpretations offered for the relationship between domain-specific nursing knowledge and the passage factors, this significant relation can also be interpreted in such a way that nursing students inevitably drew on their domain-general content knowledge to comprehend nursing English texts. On the surface our findings about the relationship between content knowledge and LSP reading performance is similar to what has been revealed in previous studies (e.g., Clapham, 1996; Krekeler, 2006; Usó-Juan, 2006), in that a significant relation can be established between content knowledge and LSP reading performance. Nevertheless, the significance in our study should not be understood in the same way as what was derived in the aforementioned studies (e.g., Clapham, 1996; Usó-Juan, 2006), in which the distinction between domain-general and domain-specific content knowledge was rarely made. Because of the lack of this distinction, the significant relationship identified in these studies can be seen as evidence supporting the overall association between background knowledge and LSP reading performance rather than as evidence verifying the hypothetical separability or inseparability of content knowledge from LSP reading performance.

Conclusion

To better understand the theoretical status of content knowledge in LSP ability, the field needs insights both into the desirability and feasibility of whether to account for content knowledge in LSP assessment. Bearing in mind the well-documented argumentative statements supporting the desirability of including content knowledge as a constituent of LSP ability, this study delved into the psychometric feasibility of separating content knowledge from LSP ability during reading score assignment. To this end, we distinguished two types of content knowledge (i.e., domain-general and domain-specific) and derived the bifactor-MIRT-based composite scores that can appreciate this distinction. By regressing the LSP reading measurement model represented by the composite scores on an external content knowledge measurement model constructed in the similar way as the LSP reading measurement model, we were able to find that it is psychometrically possible to separate the portion of domain-specific content knowledge effect from LSP reading score assignment, but this separation is impossible for the portion of domain-general content knowledge contained in the domain-general reading factor. Overall, we conclude that content knowledge is inseparable from LSP reading performance.

The results of this study have implications for LSP assessment research and practice in several ways. The first implication relates to conceptualization of content knowledge in LSP reading assessment. Previously, content knowledge has been taken as a general term and the issue of separability has been discussed under the "either-or" paradigm. This simplistic understanding of content knowledge is perhaps one of the reasons why there have been few empirical studies directly tackling the inseparability of content knowledge. By distinguishing the two types of content knowledge, our study showed for the first time that not all content knowledge is psychometrically inseparable (i.e., domain-specific content knowledge) and that not all content knowledge is separable (i.e., domain-general content knowledge).

Second, our study has implications for LSP reading test score assignment. As is well known, practice in LSP reading test score assignment is still dominated by an intention to eliminate all content-relevant information, though the importance of content knowledge to LSP ability has long been recognized by the field. The field still lacks a compelling scoring approach to account for the testlet factors that is consistent with the modern conceptualization of LSP ability, especially, LSP ability toward the more "specific" end of the general-specific continuum (Douglas, 2000; Knoch & Macqueen, 2016). The bifactor-MIRT-based composite scores could be used as an option to fill this gap. Our study not only demonstrated detailed procedures for deriving this type of score but also provided empirical evidence supporting the validity of using composite scores to represent LSP reading ability.

Finally, our finding regarding the inseparability of domain-general content knowledge from LSP reading performance has implications for assessing language for general purposes. As generally believed, a language test for general purposes should try to avoid content that would favor particular groups of test takers; and consequently, text content needs to be sampled from a broad range of knowledge domains to avoid this problem (Alderson & Urquhart, 1988). Regardless, underlying this belief and practice a fundamental question has remained unasked: If all texts are constructed in such a way, will this content neutrality introduce a general knowledge factor underlying all texts? This question should also prompt reflection over the dominant concept of the CLA: Is the claim that background knowledge is construct-irrelevant justifiable? Or has the field of language assessment ever been successful in avoiding to measure test-takers' background knowledge and not to incorporate this information in score assignment? Although seemingly beyond our current discussion, these reflections resonate well with Purpura's (2017) recent conceptualization of "meaning" in general language ability, a source of LSP ability and whose development would continue to illuminate our understanding of LSP ability.

Like other studies, our study is not without limitations. First, during bifactor-MIRT modeling, we had to exclude a few items not fitting well with the selected model. This might have led to some loss of substantive information. Second, the sample size was not large enough for performing more complex models (i.e., by accounting for the guessing parameter); it is hence unknown to what extent this simplification has led to inflated or deflated parameter estimation. Nevertheless, these limitations should not undermine our conclusion about the inseparability of content knowledge from LSP reading performance, given our focus was more on the initiation of a substantive discussion rather than on the meticulousness of psychometric modeling. Future studies can be designed to examine how factors, such as item numbers and sample size, may affect the accuracy of psychometric modeling results. The third limitation deals with our interpretation about the relationship between text features and domain-specific content knowledge effect. Because there were only four texts, we could only interpret our findings intuitively but not on statistical significance. The findings of the superficial comparison among domain-specific content knowledge effects across the texts, however, is still useful in revealing the relevance of some particular text features in indicating content knowledge effect.

Funding

This work was supported by the Educational Testing Service TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment; The University of Hong Kong Faculty of Education Research Fund; Assessment Systems Corporation Grants for Graduate Students in Psychological Measurements.

References

Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. Applied Measurement in Education, *7*(4), 255–278.

Alderson, J. C. (1981). Report of the discussion on testing English for specific purposes. In J. C. Alderson & A. Hughes (Eds.), Issues in language testing (pp. 123–134). London, UK: The British Council.

Alderson, J. C., & Urquhart, A. (1988). This test is unfair: I'm not an economist. In P. L. Carrell, J. Devine, & D. E. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 168–182). Cambridge, UK: Cambridge University Press.

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, *2*(2), 192–204. doi:10.1177/026553228500200207

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: Americal Psychological Association (APA).

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F. (2004). Statistical analyses for language assessment. Cambridge: Cambridge University Press.

Bachman, L. F., & Kunnan, A. J. (2005). Statistical analyses for language assessment workbook and CD ROM. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi:10.1007/BF02293801

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168. doi:10.1007/BF02294533

Byrne, B. M. (2010). *Structural equation modeling with Mplus: Basicconcepts, applications, and programming*. New York, NY, and London, UK: Routledge.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57. doi:10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335. doi:10.3102/1076998609353115

Cai, L., Du Toit, S., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling* [Computer software]. Chicago, IL: Scientific Software International.

Cai, L., Thissen, D., & Du Toit, S. (2011). *IRTPRO user's guide*. Lincolnwood, IL: Scientific Software International.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221–248. doi:10.1037/a0023350

Cai, Y. (2015). The value of using test responses data for content validity: An application of the bifactor-MIRT to a nursing knowledge test. *Nurse Education Today*, *35*(2), 1181–1185. doi:10.1016/j.nedt.2015.05.014

Chapelle, C., Enright, M., & Jamieson, J. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. doi:10.3102/10769986022003265

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230. doi:10.1177/1094428114555994

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge, UK: Cambridge University Press.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. doi:10.1037/0021-9010.78.1.98

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, *18*(2), 133–147. doi:10.1177/026553220101800202

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*(2), 104–121. doi:10.1177/0146621612437403

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354–378. doi:10.1080/15305058.2013.799067

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, UK: Cambridge University Press.

Douglas, D. (2013). ESP and assessment. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 367–383). Malden, MA: Wiley-Blackwell.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY, and London,UK: The Guilford Press.

Faber, P. (2012). *A cognitive linguistics view of terminology and specialized language*. Berlin, Germany: Walter de Gruyter.

Fulcher, G. (2000). The "communicative" legacy in language testing. *System*, *28*(4), 483–497. doi:10.1016/s0346-251x(00)00033-6

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436. doi:10.1007/BF02295430

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36(2), 193–202.

Huhta, M., Vogt, K., Johnson, E., Tulkki, H., & Hall, D. R. (2013). *Needs analysis for language course design: A holistic approach to ESP*. Cambridge, UK: Cambridge University Press.

IBM Corporation. (1989–2011). *IBM SPSS Statistics for Windows*, Version 20.0 [computer program]. Armonk, NY: Author.

Jöreskog, K. G., Sörbom, D., & Du Toit, S. (2001). *LISREL 8: New statistical features*. Lincolnwood, IL: Scientific Software International.

Kim, H., & Elder, C. (2015). Interrogating the construct of aviation English: Feedback from test takers in Korea. *Language Testing*, *32*(2), 1–21. doi:10.1177/0265532214544394

Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessment. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 21–38). Plymouth, UK: Rowman & Littlefield Education.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. doi:10.1037/0033-295X.85.5.363

Knoch, U., & Macqueen, S. (2016). Language assessment for the workplace. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 291–307). Boston, MA: De Gruyter Mouton.

Krashen, S., & Brown, C. L. (2007). What is academic language proficiency. *STETS Language & Communication Review*, *6*(1), 1–5.

Krekeler, C. (2006). Language for Special Academic Purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing*, *23*(1), 99–130. doi:10.1191/0265532206lt323oa

Kunnan, A. J. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1098–1114). Malden, MA: Wiley.

Kunnan, A. J. (2017). *Evaluating language assessments*. New York, NY: Routledge.

Lee, Y. W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, *21*(1), 74–100. doi:10.1191/0265532204lt260oa

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21. doi:10.1177/0146621605275414

Liu, G. Z., Chiu, W. Y., Lin, C. C., & Barrett, N. E. (2014). English for Scientific Purposes (EScP): Technology, trends, and future challenges for science education. *Journal of Science Education and Technology*, *23*(6), 827–839. doi:10.1007/s10956-014-9515-7

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First-and higher order factor models and their invariance across groups. *Psychological Bulletin*, *97*(3), 562–582. doi:10.1037/0033-2909.97.3.562

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*(2), 99–114. doi:10.1177/01466210022031552

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

METS. (2007). *Medical english test system level two (for Nurses)*. Shanghai, China: Higher Education Press.

Ministry of Health. (2007). *China nurse licensing examination syllabus 2007*. Beijing, China: People's Medical Publishing House.

Muthén, L. K., & Muthén, B. Q. (1998–2015). *Mplus 7.4* [Computer software]. Los Angeles, CA: Muthén & Muthén.

Paap, M. C., & Veldkamp, B. P. (2012). Minimizing the testlet effect: Identifying critical testlet features by means of tree-based regression. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in Practice at RCEC*. Enschede, the Netherlands: RCEC, Cito/University of Twente.

Purpura, J. E. (1998). The development and construct validation of an instrument designed to investigate selected cognitive background characteristics of test-takers. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 111–140). Mahwah, NJ: Lawrance Eribaum.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Purpura, J. E. (2017). Assessing meaning. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 33–61). New York, NY: Springer.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401–412. doi:10.1177/014662168500900409

Reckase, M. D. (2009). *Multidimensional item response theory*. London, UK, and New York, NY: Springer.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. doi:10.1080/00273171.2012.715555

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247. doi:10.1111/jedm.1991.28.issue-3

Steinberg, L., & Thissen, D. (2013). Item response theory. In J. S. Comer & P. C. Kendall (Eds.), *The Oxford handbook of research strategies for clinical psychology* (pp. 336–373). Oxford, UK: Oxford University Press.

Tapiero, I. (2007). *Situation models and levels of coherence: Toward a definition of comprehension*. London, UK: Taylor & Francis.

Thissen, D. (2013). Using the testlet response model as a shortcut to multidimensional item response theory subscore computation. In R. E. Millsap, L. A. Van Der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 29–40). New York, NY: Springer.

Usó-Juan, E. (2006). The compensatory nature of discipline-related knowledge and English-language proficiency in reading English for Academic purposes. *The Modern Language Journal*, *90*(2), 210–227. doi:10.1111/modl.2006.90.issue-2

Van Der Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: Theory and practice*. New York, NY: Springer.

Van Der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge, UK: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201. doi:10.1111/jedm.1987.24.issue-3

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203–220. doi:10.1111/jedm.2000.37.issue-3

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. doi:10.1111/jedm.1993.30.issue-3

Appendix : Technical Introduction to Bifactor-MIRT

According to this structure, the relationship between the probability of correct response to an item, given the primary (or general) dimension of ability and additional secondary factor and item characteristics, can be formulated as:

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

where θ_0 is the primary dimension of ability (general factor), θ_s the secondary factor (group factor), cthe guessing parameter (lower asymptote), d the item intercept, $a_0$the discrimination parameter on the general factor, and a_s the discrimination parameter for specific factors Cai, L., du Toit, S. H. C., & Thissen, D. (2011). To obtain stable parameter estimation for a MIRT modeling, it is necessary to have a sample size of 1,000 (Reckase, 2009) to 2000 (Ackerman, 1994) or over. Otherwise, the guessing parameter (or even the discrimination parameter) would need to be constrained (Reckase, 2009). Drawing on Reckase (2009), the multidimensional conceptualization of item difficulty (MDIFF) can be obtained by using the following equation:

$$\mathrm{MDIFF_i}=\mathrm{B_i}=\frac{-\mathrm{d}}{\sqrt{\mathrm{a_0}^2+\mathrm{a_s}^2}},$$

and the multidimensional discrimination (MDISC) can be obtained by:

$$\mathrm{MDISC}=\mathrm{A_i}=-\mathrm{d}/\mathrm{B_i}$$