วารสารอิเล็กทรอนิกส์
ทางการศึกษา

OJED, Vol.9, No.1, 2014, pp. 442-456

OJED
An Online Journal
of Education
http://www.edu.chula.ac.th/ojed

การประยุกต์ใช้วิธีการอ้างเหตุผลโดยอิงโมเดลราชในการตรวจสอบความตรงของแบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการสำหรับนิสิตระดับบัณฑิตศึกษาที่เรียนภาษาอังกฤษในฐานะภาษาต่างประเทศ

APPLYING A RASCH-BASED ARGUMENT APPROACH TO THE VALIDATION OF THE ACADEMIC COLLOCATIONAL COMPETENCE TEST FOR EFL GRADUATE STUDENTS

นายอภิชาติ คำบุญเรือง *
Apichat Khamboonruang
ผู้ช่วยศาสตราจารย์ ดร. จิรดา วุฑฒยากร **
Asst. Prof. Jirada Wudthayagorn, Ph.D.

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อประยุกต์วิธีการอ้างเหตุผลโดยอิงโมเดลราชในการตรวจสอบความตรงขั้นต้นของแบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการ แบบสอบนี้พัฒนาขึ้นโดยใช้คำปรากฏร่วมกริยาและนามที่มีความถี่สูงจากภาษาเขียนเชิงวิชาการในหลายสาขาที่อยู่ในคลังข้อความ British National Corpus แบบสอบนี้สร้างขึ้นเพื่อเป็นแบบสอบจัดระดับแบบอิงกลุ่มเพื่อวัดสามัตถิยะคำปรากฏร่วมเชิงรับของนิสิตระดับบัณฑิตศึกษาที่เรียนภาษาอังกฤษในฐานะภาษาต่างประเทศ กลุ่มตัวอย่าง คือ นิสิตระดับบัณฑิตศึกษาจำนวน 193 คน จากจุฬาลงกรณ์มหาวิทยาลัย การวิเคราะห์ข้อมูลใช้โมเดลราช ผลการวิเคราะห์พบว่า วิธีการวัดโมเดลราชให้ข้อมูลเชิงประจักษ์ที่เพียงพอและน่าเชื่อถือในการสนับสนุนเหตุผลความตรงของแบบสอบ การแปลผลคะแนนของแบบสอบมีความสมเหตุสมผลเนื่องจากหลักฐานที่ได้จากโมเดลราช ได้แก่ ความเป็นเอกมิติ ความเที่ยงของข้อสอบ (0.96) ความเที่ยงของผู้สอบ (0.86) และ การประมาณค่าพารามิเตอร์ของข้อคำถาม

* นิสิตมหาบัณฑิตสาขาวิชาภาษาอังกฤษเป็นภาษานานาชาติ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
E-mail Address: noom_linguamusica@hotmail.com
** อาจารย์ประจำสาขาวิชาภาษาอังกฤษเป็นภาษานานาชาติ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
E-mail Address: wudthayagorn@hotmail.com
ISSN 1905-4491

Abstract

The purpose of this study was to apply a Rasch-based argument approach to build the preliminary validity argument for the Academic Collocational Competence Test (ACCT). The ACCT was developed using high-frequency verb-noun collocations from varying domains of the academic written discourse in the British National Corpus (BNC) and designed primarily as a norm-referenced placement test of receptive collocational competence of EFL graduate students. A total of 193 students at Chulalongkorn University, Thailand, participated in this study. Several The data were analysed using a Rasch model. Results revealed that a Rasch measurement approach provided sound and sufficient empirical evidence in support of the validity argument for the ACCT. The ACCT score interpretation was reasonably substantiated by Rasch evidence related to unidimensionality, item reliability (0.96), person reliability (0.86), and item parameter estimation.

คำสำคัญ: วิธีการอ้างเหตุผล/วิธีการวัดแบบโมเดลราช/แบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการ

KEYWORDS: ARGUMENT-BASED APPROACH/RASCH MEASUREMENT APPROACH/ACADEMIC COLLOCATIONAL COMPETENCE TEST

Introduction

It has long been recognised that collocation plays a significant role in second language development, for it helps L2 learners use English in a more natural and accurate way (Benson, Benson, & Ilson, 2009; Hoey, 2005; Howarth, 1998; Lewis, 2000; Nation, 2001; Read, 2000; Schmitt, 2000; Sinclair, 1991). In particular, graduate students are encouraged to pay close attention to academic verb–noun collocations which are commonly found in the academic discourse and are problematic for EFL learners to properly produce (Laufer & Waldman, 2011; Ganji, 2012; Luzón Marco, 2011; Nesselhauf, 2005). As such, if teachers know to what extent students possess academic collocational ability, this may in turn help them infer to what degree students are proficient in English and who should or should not take more English courses in order to survive their advanced studies in university or other higher-education settings where English is a tool for learning.

To make proper decision as such, teachers need to rely hugely on sound and sufficient information provided by a well-developed and validated collocation test. It is thus of crucial importance that a measure of academic collocational ability be properly developed and validated to provide scores that can be meaningfully interpreted as reflecting academic collocational ability and appropriately used to facilitate teachers' decision about placement or diagnostic purposes. Proper score interpretation and use can indeed be highly beneficial for test-takers and test users alike, whereas misinterpretation or misuse of test scores might go the other way round. While a multitude of research in the literature has thus far been conducted specifically to develop and validate vocabulary tests, far less research (e.g., Jaen, 2007; Keshavarz & Salimi, 2007; Voss, 2012) has been done to develop measures of collocation knowledge especially using advanced measurement methods. Only relatively recently has there been a few studies (e.g., Voss, 2012) set out to

develop and validate collocation test using a corpus-based approach, a Rasch psychometric approach and an argument-based validation framework.

What we have rationalised previously essentially underpins the primary objectives of the present study. This study aimed to develop the ACCT and build the validity argument for the ACCT using a Rasch measurement approach under the framework of Kane's argument-based approach. The hybrid of two scientific models was of greater help to validate the appropriateness of the score interpretation and use of the ACCT, which was developed using a five-option multiple-choice format, based on a corpus-driven method, and designed primarily as a norm-referenced placement test of students' receptive collocational competence.

An argument-based approach to validation

From the perspective of an argument-based approach (Kane, 1992, 2013), validation is to validate test score interpretation and use by evaluating the feasibility of the proposed interpretation and use of test scores. Therefore, the proposed interpretation and use of test scores need to be initiated as clearly as possible. Kane's argument-based approach involves two argument development stages. The first stage is to develop the interpretive/use argument by specifying the intended interpretation and use of test scores. The second step is to build the validity argument by evaluating a priori and empirical evidence sought to support such intended interpretation and use of test scores outlined in the interpretive/use argument. In this study, the ACCT interpretive argument focused on five inferences in the TOEFL interpretive argument (Chapelle, 2008; Chapelle, Enright, & Jamieson, 2010): domain description, evaluation, generalisation, explanation, and extrapolation. Each inference has its warrant which rests on assumptions requiring empirical backing from a Rasch measurement analysis.

In the ACCT interpretive argument, the domain description inference warrants that student performances on the ACCT reveal the collocational competence relevant to and representative of the target language use (TLU) domain in university or other higher-education settings. This warrant assumes that: 1) collocations on the ACCT are representative of the TLU domain of the academic written discourse, and 2) the ACCT can elicit student responses which reflect the collocational competence. The evaluation inference has a warrant that observed responses on the ACCT are evaluated to provide observed scores reflective of the collocational competence. This warrant rests on assumptions that scoring procedure is appropriate to elicit student responses which serve as evidence of varying levels of the collocational competence.

The generalisation inference warrants that observed scores on the ACCT are estimates of expected scores which are congruent across items and invariant across gender. This warrant assumes that: 1) estimates of studentperformance can consistently distinguish

among students, and 2) item estimates are invariant across gender. The explanation inference warrants that expected scores are attributed to the collocational competence comstruct in the academic written discourse. This warrant assumes that: 1) performance on the ACCT reflects students' collocational competence, and 2) student responses to distractors on the ACCT are consistent with the intended cognitive process around which distractors are developed. The extrapolation inference warrants that the collocational competence construct as measured by the ACCT accounts for relevant language performance in the academic discourse in university or other higher-education settings. This warrant assumes that the ACCT scores can distinguish students with different levels of English proficiency.

It is important to realise, however, that gathering all evidentiary information to support validity is a lengthy or even endless process, depending on how sophisticated the proposed score interpretation and use are (Kane, 2013). In this study, the validity argument was based on an evaluation of five inferences: domain description, evaluation, generalisation, explanation, and extrapolation. This study aimed to map several applications of a Rasch measurement approach onto Kane's argument-based validation framework with a view to providing preliminary empirical evidence in support of the ACCT validity argument.

A Rasch measurement approach to validation

To provide preliminary evidence for the ACCT validity argument, a Rasch model analysis was used in this study to investigate psychometric properties of the ACCT which can serve as preliminary validity evidence. Unlike CTT, a Rasch model has a major advantage over CTT in that it uses mathematical models to predict probability estimates for both person ability and item difficulty that are independent of a particular group of examinees or a set of items (Bond & Fox, 2007; Embretson & Reise, 2000; Rasch, 1980). A Rasch model offers several applications that can be used to provide empirical evidence supporting the inferences in the ACCT interpretive argument.

In the domain description inference, the point-measure correlation can be used to check the adequacy of item content and the congruency of a particular item with the remaining items on the instrument. The correlation should be positive to show the correlation between scores on the item and scores on the remaining items. The value close to zero means that items are too easy or difficult to answer correctly or they do not measure the construct in the same manner as other items do (Wolfe & Smith, 2007). The item fit indices can be used to investigate the unidimensionality of the items or other measurement problems. Item fit indices indicate whether the test content is relevant to the intended construct and assure that items elicit a relevant, unidimensional construct of interest, while misfit items may assess irrelevant, subdimensional constructs (Wolfe & Smith, 2007).

As for the evaluation inference, the principal component analysis of linearized Rasch residuals (PCAR) can be used to check the unidimensionality of the data by determining whether there is a sufficient amount of variance explained by the construct in question. If the data fit the model, it can then be confident that item scoring is appropriate for eliciting the construct under measure (Wolfe & Smith, 2007). As for scoring, a Rasch dichotomous model scales observed scores into comparable measured scores, hence contributing to the standardisation of scoring process (Aryadoust, 2009; Schumaker, 2004; Wolfe & Smith, 2007). Transforming raw scores to measured scores in the Rasch analysis is of fundamental importance, for the distance between measured scores is equal and thereby item difficulties can be compared with person abilities (Bond & Fox 2007; Rasch, 1980; Wolfe & Smith, 2007).

In respect of the generalisation inference, Wolfe and Smith (2007) suggest that the item reliability informs how well examinee abilities spread out items difficulties or how well item difficulties are dispersed along the difficulty hierarchy. The item separation supplements the item reliability by checking how well items are classified into different levels on the item difficulty hierarchy. Another useful index is the item strata index which indicates whether person competencies statistically distinguish item difficulty levels. The person reliability (analogous to coefficient alpha and KR-20) can be employed to check how well item difficulties spread out examinee abilities or how well competencies are distributed along the competence hierarchy. The person separation supplements the person reliability by examining to what extent persons are separated into different competency levels on the competency hierarchy. The person strata index also indicates how well items statistically discriminate competence levels.

With respect to the explanation inference, Linacre (2012) and Wolfe and Smith (2007) suggest that the item fit statistics and PCAR give useful information on the relevancy and unidimensionality of the construct being measured. Regarding the extrapolation inference, the person strata index can be used to inform how well items statistically classify person abilities. The person strata index greater than 2 suffices to confirm that items distinguish the more competent from the less competent. Although a Rasch model has long taken its place in language testing (McNamara & Knoch, 2012), only a few collocation tests have been validated using a Rasch model (Voss, 2012) while much more vocabulary tests has been evaluated using a Rasch model and Messick' validity framework (e.g., Beglar, 2010; Baghaei & Amrahi, 2011).

Objective

The primary purposes of this study were to develop the ACCT for EFL graduate students and to apply a Rasch measurement approach to provide evidence in support of

the ACCT validity argument under the framework of Kane's argument-based approach which focuses on validating the appropriateness of the interpretation and use of the ACCT scores.

Methodology

Participants

The participants were 193 graduate students with different levels of English proficiency and from varying disciplines at Chulalongkorn University, Thailand. The students were grouped into high, moderate, and low proficiency levels based on Chulalongkorn University Test of English Proficiency (CU-TEP), TOEFL iBT, and IELTS scores they used to apply for the university. The criteria for classifying proficiency groups were presented in Table 1 and demographic characteristics of students are presented in Table 2.

Table 1. Criteria for classifying proficiency groups

English Proficiency	CU-TEP	TOEFL iBT	IELTS
Low	0 - 449	0 - 44	0.0 - 4.5
Moderate	450 - 579	45 - 91	5.0 - 6.0
High	580 - 677	92 - 120	6.5 - 9.0

Table 2. Demographic characteristics of 193 EFL graduate students

Demographic Characteristics	Proficiency level							
	Low		Mid		High		Total	
	n	%	*n*	%	*n*	%	*n*	%
Gender								
Male	32	49.2	22	33.8	11	16.9	65	33.7
Female	52	40.6	37	28.9	39	30.5	128	66.3
Study level								
Master	84	49.1	49	28.7	38	22.2	171	88.6
Doctor	0	0.0	10	45.5	12	54.5	22	11.4
Native language								
Thai	82	46.3	52	29.4	43	24.3	177	91.7
Chinese	1	12.5	3	37.5	4	50.0	8	4.1
Vietnamese	0	0.0	2	50.0	2	50.0	4	2.1
Lao	1	50.0	1	50.0	0	0.0	2	1.0
Hindi	0	0.0	0	0.0	1	100	1	0.5
Cambodian	0	0.0	1	100.0	0	0.0	1	0.5
Total	84	43.5	59	30.6	50	25.9	193	100

Instrument

The ACCT is a paper-delivered multiple-choice test and was developed by the authors to measure the ability to recognise verb-noun collocations used in academic written English. It was developed based on high-frequency verb-noun collocations from BNC.

Colocations and test inputs were extracted through the Lancaster BNCweb Server. A five-option multiple-choice item format was chosen for the current collocation test as it is appropriate to measure a receptive collocational competence. The initial version of the ACCT contained 50 items.

In the pilot study, 50 items were administered to 30 graduate students with low, moderate and high English proficiency and were then analysed using the TAP item analysis software. Items that had difficulty index of between 0.2 and 0.9 and discrimination index at least 0.2 were included to compose the final ACCT. After piloting and evaluating 50 ACCT items, the final version of the ACCT consists of 30 items with the Cronbach's alpha internal consistency coefficient of 0.85. The time allowed for the 30-item ACCT is 30 minutes. Figure 1 shows an example of the ACCT Item 21. Test questions are incomplete sentences. Beneath each sentence, there are five verbs, marked a, b, c, d, and e. Examinees had to choose one verb that best collocates with a noun in the sentence with the most appropriate meaning for the academic context.

21) In 1986, as part of its wider proposals for the reform of local government finance, the government declared its intention to _____ a new grant **system.**
a. renovate b. integrate c. introduce d. invent e. install

Figure 1. An example of an ACCT item

Procedure

The ACCT was administered to graduate students from December 2013 to January 2014, together with the 30-item Academic Vocabulary Levels Test Version 2 (Schmitt, Schmitt, & Clapham, 2001) and the 3-item test reflection questionnaire. Time allowed for the tests was 60 minutes. The tests were counterbalanced and administered during certain class periods. We asked for approval from teachers responsible for the classes and asked for cooperation from volunteer students. The primary author delivered the ACCT, explained the test instruction both in Thai and in English, and monitored students.

Results and discussion

Descriptive statistics

Descriptive statistics in Table 3 showed that the ACCT scores for 193 graduate students were normally distributed. The mean score for 30 items was 14 with the standard deviation of 6.85. The skewness and kurtosis values were 0.37 and -1.091 respectively, both indicating a normal distribution of the ACCT score data. As for subgroups, the scores of low, moderate and high groups were also normally distributed. The mean scores of low, moderate and high groups were 8.47, 15.06, and 22.56 respectively and the standard deviation of low, moderate and high groups were 2.69, 5.13, and 3.49 respectively. The skewness and kurtosis values of all groups were in the range of between +2 and -2, thereby indicating normal distributions of subgroup scores.

Table 3. Descriptive statistics of the ACCT scores

Group	N	M	S.D.	Range	Min	Max	SK	KU
Low	84	8.47	2.69	13.00	3.00	16.00	.23	-.39
Moderate	59	15.06	5.13	21.00	4.00	25.00	-.05	-.82
High	50	22.56	3.49	15.00	14.00	29.00	-.28	-.12
Total	193	14.13	6.85	26.00	3.00	29.00	.37	-1.09

Unidimensionality

The data were analysed using the Winsteps software (version 3.80.1). The PCAR showed that the amount of the variance explained by different components in the data was 32% with 14.9% explained by persons and 17.1% explained by items. The unexplained variance of the first contrast was 6.1 with the eigenvalue of 2.7. Reckase (1979) suggests that the variance explained by the focal factor should be greater than 20% to ensure the unidimensional construct. Linacre (2012) recommends that the unexplained variance of the first contrast should not exceed 5% and the first contrast eigenvalue should not exceed 3.

Since the variance of the focal collocational construct was explained by more than 20% and the first contrast eigenvalue was less than 3, we assumed that the focal collocational construct was substantively unidimensional. This was also substantiated by the evidence that 29 items possessed good fit indices and had positive point-measure correlations. Overall, the PCAR and item fit statistics signified the substantive unidimensionality of the collocational construct.

Internal consistency reliability

For 30 ACCT item and 193 students, the item reliability was 0.96, indicating that students well spread ACCT item difficulties or item difficulties were widely dispersed on the difficulty hierarchy. The item separation was 4.90, indicating that ACCT items were separated into around five difficulty categories. In other words, students statistically differentiated more difficulty items from easier items. The item strata was 6.86, meaning that student competencies statistically distinguished approximately six item difficulty levels. This suffices to say that the ACCT contained adequate items to reliability measure students.

The person reliability was 0.86 and the coefficient alpha of .89, meaning that ACCT items well differentiated students in terms of collocational competency or students' collocational competencies were well dispersed on collocational competence hierarchy. The person separation was 2.48, indicating that student competency was classified into roughly two groups on the collocational competency scale. In other words, the ACCT items statistically distinguished higher-ability persons from lower-ability students. The person strata index was 3.64, demonstrating that the ACCT items statistically differentiated approximately three collocational competence levels.

Item parameter estimation

Table 4 shows item statistics of 30 ACCT items. Item difficulty ranged between 1.75 (Item 2) and -2.14 measures (Item 3). The mean difficulty was 0 and the standard deviation was 0.91, indicating that the ACCT difficulty was average. Linacre (2012) suggests including any items with a Mnsq value of between 0.5 and 1.5 for productive items. Items 19, 2, 13, and 28 had Outfit Mnsq values greater than 1.5 and hence appeared underfit to the Rasch model. Item 19 was the most underfit (Outfit Mnsq = 2) and therefore was first deleted prior to reanalysing the new data set. After reanalysing the new data set, Item 2, 13, 28, remained underfit and Item 23 turned out underfit to the Rasch model. However, Infit Mnsq values of Items 2, 13, 23, and 28 fell within 0.5 and 1.5 and their Outfit Mnsq values were slightly beyond 1.5. If these items were deleted, the remaining items may not well represent the collocational competence construct (Linacre 2012). On this account, we decided to keep these items on the ACCT.

Table 4. Item measure statistics of 30 ACCT items

Item	Collocation	Total Score	Item Difficulty	Model S.E	Infit Mnsq	Outfit Mnsq	PT Measure
01	find a way	106	-0.44	0.17	1.16	1.21	.37
02	cite a case	37	1.75	0.21	1.18	1.55	.31
03	leave school	160	-2.14	0.20	0.95	0.78	.38
04	enforce a law	100	-0.27	0.17	0.76	0.68	.66
05	make an award	96	-0.16	0.17	1.13	1.13	.41
06	cover an area	116	-0.71	0.17	0.90	0.77	.55
07	see figure	92	-0.05	0.17	0.65	0.59	.73
08	obtain a result	103	-0.36	0.17	1.11	1.06	.42
09	provide an example	117	-0.74	0.17	1.02	1.12	.43
10	improve health	113	-0.63	0.17	1.08	1.20	.41
11	conduct a study	96	-0.16	0.17	0.61	0.54	.76
12	have an idea	71	0.55	0.17	0.89	0.85	.58
13	make sense	156	-1.98	0.20	1.15	1.56	.19
14	justify belief	111	-0.58	0.17	0.94	0.87	.52
15	hold the view	61	0.86	0.18	0.95	0.97	.53
16	account for the fact	57	1.00	0.18	0.92	0.90	.55
17	play a part	56	1.03	0.18	0.92	0.84	.56
18	pursue a policy	56	1.03	0.18	1.00	0.99	.49
19	fight the war	89	0.03	0.17	1.70	2.00	.02
20	exercise power	51	1.20	0.19	0.89	1.13	.53
21	introduce a system	49	1.27	0.19	0.97	0.99	.49
22	apply a rule	103	-0.36	0.17	0.98	0.94	.51
23	carry on a business	85	0.15	0.17	1.38	1.43	.26
24	appoint an expert	89	0.03	0.17	0.73	0.70	.68
25	terminate a contract	97	-0.19	0.17	0.91	0.87	.55
26	use a word	129	-1.09	0.17	0.76	0.63	.61

Item	Collocation	Total Score	Item Difficulty	Model S.E	Infit Mnsq	Outfit Mnsq	PT Measure
27	do work	103	-0.36	0.17	0.93	0.88	.54
28	read text	45	1.42	0.20	1.20	1.59	.31
29	have a disease	75	0.43	0.17	1.15	1.13	.41
30	treat a group	110	-0.55	0.17	0.98	0.92	.50
	Mean	91.0	0.00	0.18	1.00	1.03	
	S.D	30.2	0.91	0.01	0.21	0.33	

Overall, 29 items exhibited good fit and had a non-zero positive point-biserial correlation, thus well fit the expected Rasch model. Only Items 19 was critically underfit to the Rasch model due probably to its tack of unidimensionality or unexpected variance related to guessing or carelessness (Linacre, 2012; Wolfe & Smith, 2007). Figure 2 shows a person-item babble map presenting the precision and accuracy of the person ability estimate and item difficulty estimate. The precision of the estimate can be examined through standard error of measurement, while the accuracy of the estimate is examined through the model fit.

A person-item babble map alights each person ability in darker colour and item difficulty in lighter colour vertically onto the same standardised interval scale, logit, or measure which has equal distances or units and ranges from +5 at the top, 0 in the middle, and down to -4 at the bottom. Higher positive values represent more competent persons and more difficult items, whereas lower negative values represent less competent persons and less difficult items. The measurement error of both person and item estimates is expressed by the size of the symbol, the larger the symbols, the greater the errors, and hence the lower the precision of the estimates.

The accuracy of person and item estimates is expressed in terms of how far items and persons are from the acceptable Outfit Mnsq zone on the horizontal axis. The farther the symbols from the acceptable Outfit Mnsq zone, the lesser the model fit, and hence the lower the accuracy of the estimates. Items and persons are horizontally located onto the standardized scale, ranging roughly between +4 and -4. Items and persons that acceptably fit the expected Rasch model are located within the Outfit Mnsq zone of between 0.5 and 1.5. Items falling outside of this zone on the left are considered as overfitting items, indicating that the responses are too predictable, whereas items falling outside of this zone on the right are considered as underfitting items, indicating that responses to these items are too unpredictable. As displayed in Figure 2, only Item 19 is underfitting to the model, meaning that responses to the item are too unpredictable and may measure some related sub-dimensions that are irrelevant to the focal construct of the collocational competence. This indicates an indication of construct-irrelevant variance. For precise measurement, item

difficulties should measure a single unidimensional construct, and spread out widely on the item difficulty hierarchy.

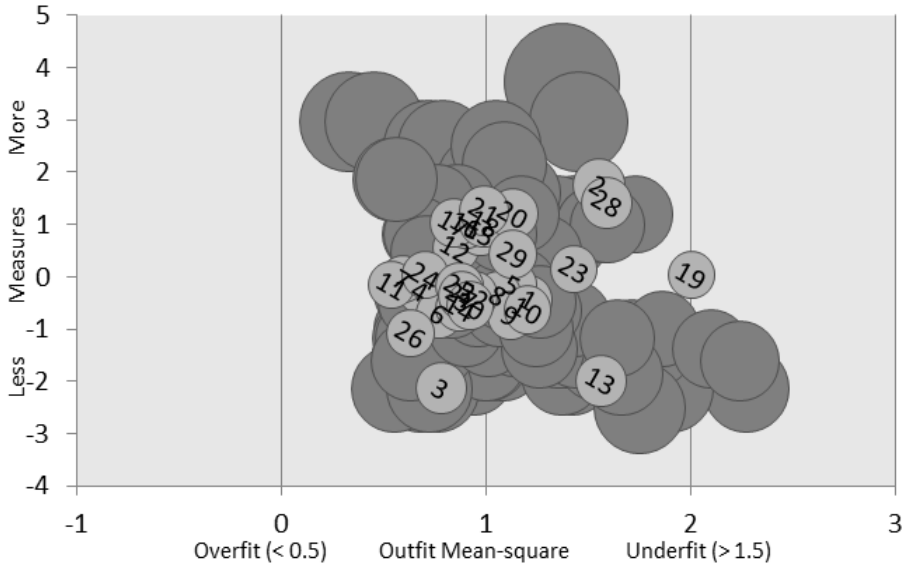

Figure 2. A person-item babble map based on Outfit Mnsq

Conclusions and suggestions

The ACCT was developed as a measure of receptive collocational competence which may be used to inform educational decision on placing students into appropriate levels in university or other higher-education setting. The ACCT was developed using high-frequency verb-noun collocations, systematically selected from seven academic domains in the academic written prose embedded in BNC. A Rasch-based Kane argument approach was used to seek empirical evidence reinforcing the ACCT validity argument. The interpretive argument focused on five inferences, each of which rests on a warrant based on underlying assumptions that necessitate sound and sufficient evidential backing from a Rasch-based data analysis. It was found that a Rasch measurement approach provided reasonable evidence supporting the ACCT validity argument.

Assumptions underlying the domain description inference were properly supported by Rasch indices. The item strata index indicated that ACCT items were categorised into around six difficulty levels and the point-measure correlation values were over zero and positive. All these pieces of evidence reasonably ensured that collocations on the ACCT were representative of the TLU domain of the academic written discourse. Moreover, the item fit statistics showed that 29 items well fit the Rasch model, meaning that the ACCT can elicit student responses which reflect the collocational competence. The assumption behind the evaluation inference was satisfactorily supported by Rasch evidence. The Rasch dichotomous model scaled observed scores into comparable, interval data, hence contributing to the standardization of scoring process. Furthermore, the PCAR indicated a

dominant unidimensional collocational construct and hence the scoring procedure was appropriate for eliciting the collocational competence construct.

With regard to the generalisation inference, the underlying assumptions were properly substantiated by Rasch evidence. Item reliability, separation, and strata and person reliability, separation, and strata were beyond acceptable criteria. Assumptions underlying the explanation inference were reasonably supported by Rasch indices. The PCAR indicated a dominant unidimensinal collocational construct measured by ACCT items. In terms of the extrapolation inference, the underlying assumption was well supported by the Rasch evidence. The person strata index revealed that about three distinct competency levels were differentiated by ACCT items.

It is evident that a Rasch measurement approach provides sound and sufficient evidence strengthening the ACCT validity argument. Rasch indices and visual plots empirically serve as essential psychometric properties of the ACCT, including unidimensionality and local independence, internal consistency reliability, and item fit indices. These psychometric properties are considered as empirical evidence supporting the ACCT validity argument. This study highlights the cost-effective, time-saving advantages that a Rasch measurement approach offers to test developers and validation frameworks, particularly Kane's argument-based approach. Table 5 summarises Rasch evidence in support of the ACCT validity argument.

Table 5. Summary of Rasch evidence in support of the ACCT validity argument

Inferences	Warrants	Assumptions	Rasch evidence
Domain description	Student performances on the ACCT reveal the collocational competence relevant to and representative of the TLU domain in university or other higher-education settings.	1) Collocations on the ACCT are representative of the TLU domain of the academic written discourse. 2) The ACCT can elicit students' responses which reflect the collocational competence.	- Rasch item fit indices - Rasch item strata - Rasch point-measure correlation
Evaluation	Observed responses on the ACCT are evaluated to provide observed scores reflective of the collocational competence.	1) Scoring procedure is proper to elicit student responses which serve as evidence of varying levels of the collocational competence.	- Principal component analysis of linearized Rasch residuals - Rasch dichotomous model
Generalization	Observed scores on the ACCT are estimates of expected scores which are congruent across items and invariant across gender.	1) Estimates of student performance can consistently distinguish among students. 2) Item estimates are invariant across gender.	- Rasch internal consistency reliability indices
Explanation	Expected scores are	1) Performances on the ACCT	- Rasch item fit indices

Inferences	Warrants	Assumptions	Rasch evidence
	attributed to the collocational competence construct in the academic written discourse.	reflect students' collocational competence. 2) Student responses to distractors on the ACCT are consistent with the intended cognitive process around which distractors are developed.	- Principal component analysis of linearized Rasch residuals
Extrapolation	The collocational competence construct as measured by the ACCT accounts for relevant language performance in the academic discourse in university or other higher-education settings.	1) The ACCT scores can distinguish students with different levels of English proficiency.	- Rasch person strata

This study carries significant implications. The current findings will inform test developers of deploying the hybrid of a Rasch model and an argument-based model to validate the score interpretation and use of language assessment instruments. This study will also exemplify a way of assessing specific collocational knowledge as an indicator of general English proficiency and as a construct of a measure for placement decision in academic English courses in university or other higher-education settings. Pedagogically, the findings could raise the awareness of teaching and learning English collocations in English classroom, which will in turn lead to positive washback. However, students' cognitive response process was not sufficiently investigated since the focus of this study was on a Rasch model analysis. The utilisation inference was not sufficiently examined in the present study and further research is needed to apply a Rasch model to examine cut-score thresholds and classification consistency to support the utility and wasback of the ACCT, thereby solidifying the ACCT validity argument.

References

Aryadoust, V. (2009). Mapping Rasch-based measurement onto the argument-based validity framework. *Rasch Measurement Transactions, 23,* 1192–1193.

Baghaei, P., & Amrahi, N. (2011). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2,* 052–1060. doi:10.4304/jltr.2.5.1052-1060

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27,* 101–118. doi: 10.1177/0265532209340194

Benson, M., Benson, E., & Ilson, R. (2009). *The BBI combinatory dictionary of English*. Philadelphia, PA: Benjamins.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. E. Enright & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). New York, NY: Routledge.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference?. *Educational Measurement: Issue and Practice, 29,* 3–13.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Ganji, M. (2012). On the effect of gender and years of instruction on Iranian EFL learners' collocational competence. *Canadian Center of Science and Education, 5,* 123–133. doi:10.5539/elt.v5n2p123

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Howarth, P. A. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19,* 24–44.

Jaen, M. M. (2007). A corpus-driven design of the test for assessing the ESL collocational competence of university students. *International Journal of English Studies, 7,* 127–147.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

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

Keshavarz, M. H., & Salimi, H. (2007). Collocational competence and cloze test performance: A study of Iranian EFL learners. *International Journal of Applied Linguistics, 17,* 81–92.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61,* 647–672. doi: 10.1111/j.1467-9922.2010.00621.x

Lewis, M. (2000). *Teaching collocation: Further developments in the lexical approach*. Hove: Language Teaching Publications.

Linacre, J. M. (2012). *A user's guide to Winsteps ministep Rasch-model computer programs*. Retrieved from http://www.winsteps.com/a/winsteps-manual.pdf

Luzón Marco, M. J. (2011). Exploring atypical verb+noun combinations in learner technical writing. *International Journal of English Studies, 11,* 77–95.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29,* 555–576. doi: 10.1177/0265532211430367

Nation, P. (2001). *Learning vocabulary in another language*. New York: Cambridge University Press.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Rasch, G. (1980). *Probabilistic model for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18,* 55–88.

Schumaker, R. E. (2004). Rasch measurement: The dichotomous model. *Journal of Applied Measurement, 4,* 87–100.

Sinclair, J. (1991). *Corpus, concordance, and collocation*. Oxford: Oxford University Press.

Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3539432)

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement, 8,* 204–234.