Journal of English Language
Teaching and Learning
University of Tabriz
No. 22, 2018

Establishing an Argument-Based Validity Approach for a Low-Stake Test of Collocational Behavior *

Ali Darabi Bazvand**
Assistant Professor, University of Human Development, College of Languages, English Department, Kurdistan, Iraq (corresponding Author)
Alireza Khorram
Assistant Professor, Shahid Chamran University of Ahvaz, Ahvaz, Iran
Seyyed Ali Mirsalari
Assistant Professor, Faculty of Humanities, Islamic Azad University of Ramhormoz, Ramhormoz, Iran

Abstract
Most of the validation studies conducted across varying test application contexts are usually framed within the traditional conceptualization of validity and therefore lack a comprehensive framework to focus on test score interpretations and test score use. This study aimed at developing and validating a collocational behavior test (CBT), drawing on Kane's argument-based approach to validity. Four types of inferences including observation, generalization, extrapolation and explanation were articulated. To verify the validity assumptions, both theoretical and empirical evidence were presented within the formative and summative stages of test development and validation. Followed from Kane, theoretical evidence was sought through test specification, item construction, and test construction procedures. Empirical support, however, was sought through examining the collocational behavior test (CBT) with a sample of 60 university students majoring in TEFL. Ebel's criteria, KR-21 reliability and a series of Pearson-Product correlation were applied to analyze the data for both theoretical and empirical phases. The findings refer to the support for the assumptions proposed for test validity, suggesting that the collocational behavior test(CBT)may provide an appropriate and accurate indicator of collocational language ability for EFL learners. The implications for language testing and assessment are discussed.
Keywords: collocational behavior, validity argument, theoretical evidence, empirical evidence

*Received date: 2018/08/26 Accepted date: 2018/12/01
**E-mail: alidarabi1350@gmail.com

Background

It is widely accepted that L2 learners may have lots of problems when dealing with vocabulary learning. The problems of appropriate lexical choice and lack of a good distinction of near synonyms are among the semantic handicaps which are daunting for EFL learners. Even advanced language learners may have difficulty handling semantic or lexical patterns. For this reason, Jaen (2007, p.127) contends that "lexis is at the heart of language acquisition".

Being able to produce a combination of words appropriately means having *phraseological* or *collocational* ability, what Pawley and Syder (1983) suggest as indicator of native–like fluency. In collocational ability, meaning is supposed to be restricted and displayed in discourse, that is, in the company of other words (Sinclair, 1991; Stewart, 2010). By way of example, *meal* is a preferred collocate for *substantial* than for *food*, whereas *big* more appropriately co-co-occurs with *food*. Without having a good knowledge of collocation we would not know whether the single word *substantial* collocates with the word *meal* or *food* (substantial meal is preferred to substantial food). Knowledge of these word restrictions is considered as knowledge of collocational patterns (Almela, 2007; Widdowson 2007). Due to the essentiality of the collocations in language pedagogy, establishing the validity of collocation measures may seem urgent in an EFL context. To date, the validity of most of the instruments measuring EFL learners' knowledge of collocational behavior seemed to have been established via traditional models, namely content, construct or criterion. As such, validating a test of collocation refined via a modern approach to validity like evidence-centered assessment design' (Mislevy, 2007), 'assessment use argument' (Bachman, 2005; Bachman & Palmer, 2010), or 'argument - based validity'(Kane, 2006, 2007, 2010, 2013)is relevant. Kane's approach to validity is strongly supported and abundantly taken up by different researchers since it seems to address the limitations attributed to traditional approaches to validity. Therefore, thanks to its all-encompassing and illuminating nature, this argument-based model was the method of choice.

Argument- Based Validity

An argument-based approach to validation was suggested by Kane (1992, 2001, 2002) by building on the work of Cronbach (1971, 1988), House (1980), and Messick (1989). Kane (1990) presented argument-based approach to validity to address the limitations attributed to traditional approaches to validity. Some of these limitations, to paraphrase Kane (2011), are having difficulty in implementing these approaches effectively, having no clear guidance on how to proceed and finally not having a well-accepted criterion for measuring their progress. Argument-based approach to validity makes use of two kinds of arguments: an interpretive argument which clarifies the proposed interpretations and uses of the results of assessment by presenting a chain of inferences and assumptions guided from observed performance to the conclusions and decisions, and a validity argument which has the job to evaluate the plausibility of the inferences and assumptions of the interpretive argument (Kane, 1992).

Validation is a continuous process involving not only the accumulation of evidence to test score interpretation and uses (Kane, 2006) but judgment about the plausibility of those interpretations (Knoch & Elder, 2013; Xi, 2008).Therefore, validity is an argument construed by an analysis of theoretical and empirical evidence instead of a collection of separate quantitative or qualitative evidence (Bachman, 1990; Chapelle, Enright, & Jamieson., 2008, 2010; Kane, 1992, 2001, 2002; Mislevy, 2003)..

According to Kane (1990) there are two important stages involved in the processes of evaluation of interpretive argument: formative and summative. The former embarks on the explicit definition of interpretive argument together with the development of a preliminary case for the plausibility of that argument. As such, the very processes of test development including test specification, test rubric, item construction, and test construction can be categorized under the formative stage of interpretive argument. As Kane (1990) argues, we can also categorize the formative stage as represented in the theoretical evidence proposed and presented in the interpretive argument. The

latter, however, necessitates the empirical control on the inferences of and assumptions in the interpretive argument.

Argument-Based Studies

Research on argument–based validity shows that there have not been many documented studies using Kane's approach (McNamara & Roever, 2006). In line with Kane's argument-based approach was the most thorough study conducted by Chapelle et al. (2008) where providing a good evidence for the validity of the new internet-based Test of English as a Foreign Language (TOEFL IBT) was a major concern. Six types of inferences illuminated their study: domain description, evaluation, generalization, explanation, extrapolation, and utilization. This particular study changed the atmosphere of language testing from a highly abstract unified model of validity to a more transparent and usable argument-based approach to validation (Bachman, 2005; McNamara & Roever, 2006), moving the way forward and paving it for subsequent argument-based (validity) studies.

In a similar study using the same framework, Abdul Kadir (2008) evaluated the effect of an English Language Proficiency Assessment (ELPA).To examine the claims for the use of the test in an argument-based approach to validity, she applied four types of inferences to change this project into a concrete encounter: scoring, generalization, extrapolation and test impact. The findings suggest that the ELPA provides an effective indicator of English language competency, regardless of the service being assessed. However, this study suffered in the sense that it failed to include some important bridges of interpretive argument. These bridges are: domain definition, evaluation, explanation, utilization and the most important inference of ultimate actions. She used scoring inference which is just one part of evaluation inference, neglecting the characteristics of the test and the conditions of test administration as other important components of evaluation inference. Moreover, as Abdul Kadir herself pointed out, the data collected to inform generalizability inference did not rely on operational data; rather rescoring was necessary to get the data for this level. Moreover, the sample size used to inform the generalizability

inference was limited (less than 15% of the total available data). All in all, giving imbalance weight to test score interpretation and use may render this project as incomplete.

Relying on a balanced focus on both test score interpretation and test utilization, Riazi and Johnson (2013) applied a hybrid of two validation structures – Kane's interpretive model and Bachman's assessment use argument for a standardized placement test (Accuplacer), and a locally developed and marked writing sample. The study benefited from different types of evidence such as instrument outcomes, student course results, institutional practices and policies, test publisher data, and the opinions of stakeholders gathered via focus group interview and questionnaires. The core of this study was based on three technical bridges of Kane's framework including evaluation, generalization and extrapolation and the decision and consequence inferences adapted from Beckman and Palmer's argument approach.

Applying more than one single framework to investigate validity studies to comprehensively account for all the evidences of test interpretation and use, Riazi and Johnson's hybridized study is informing not only for the current study but for other future studies delving into the argument-based approaches to validity. Practically, however, no suggestion has been made regarding ultimate actions to be taken by test practitioners for test misuses. Ultimate actions are hypothesized mechanisms that are supposed to address unintended consequences. These actions are to be taken to improve the technical and decision quality of tests in different contexts

In line with Riazi and Johnson's study is Chung's (2014) investigation which was motivated to supplement the decisions made about nonnative English speaking students' placements in ESL writing courses on the basis of their performance on a test of productive grammatical writing ability in academic English (i.e., the academic grammar test). This project relied on a localized version of Chapelle et al.'s (2008, 2010) validity framework, which was developed from Kane's (2006) interpretive/validity argument model. It also adopted some aspects of Bachman and Palmer's (2010) Assessment Use

Argument (AUA) model, thus being a hybridized framework. Seven inferences (domain description, evaluation, generalization, extrapolation, explanation, utilization, and ramification) informed Chung's interpretive argument. There are some problems with this study. For example, in the domain description, the critical skills and abilities of target language use domain have not been identified and it definitely casts doubt on how Chung has evaluated the representativeness of the observed knowledge, skills and abilities (demonstrated by test takers) to those of target language use domain. Moreover, like that of Riazi and Johnson, Chung's study does not include ultimate actions as the last element of the chain of inferences in the interpretive argument.

In a nutshell, we can argue that the studies mentioned above, in one way or another, opted for an argument-based approach which as, Kane (1990) believes, is a pragmatic approach to test validation. However, none of the studies mentioned above simultaneously takes the formative- summative stages of interpretive argument to investigate validity in language components or sub-components. Park's (2012) study only follows formative-summative stage but it is focused on inferential reasoning which is related to statistics. Thus, the present study tries to address this gap.

Research Questions

This study aims to develop and validate a Collocational Behavior Test (hereafter CBT). It specifically tries to embark on four levels of inference taken from Kane's (2006) interpretive argument. Therefore, the following research questions aim to address some of the concerns raised for this study. These questions were formulated according to the levels of inference in line with the argument approach to test validation. The corresponding inference for each research question is listed in parentheses:

1. Are the procedures for CBT development detailed and based on test tasks presented in test specification? (observation)
2. To what extent the items included in the collocation test enjoy acceptable discrimination and reliability indices? (Generalization)

3. Is there a strong correlation between the collocation test and a criterion measure of collocation ability as expected? (Extrapolation)
4. Is there a strong correlation between scores on the total test and scores on the receptive and productive sub-tests? (Explanation)

Method

Framework of the Study

Within the framework of an argument-based approach to validation, this study was structured to formative and summative stages proposed by Kane (1992). In the formative stage, the test specification and detailed procedures for test development were organized and a preliminary assessment of CBT was made by expert reviews. Assumptions, warrants and backing of the observation inference were proposed and formulated in this stage. In the summative stage, the interpretive argument was examined in terms of plausibility of the associated assumptions specified by some empirical evidence represented in generalization, extrapolation, and explanation inferences. The validity of the interpretive argument was strengthened to the extent that each type of evidence supports the inferences and assumptions regarding score interpretations and uses (Kane, 2006a, 2006b).

Participants

The study involved 60 junior students majoring in TEFL. They were registered in a four-year B.A. program at Islamic Azad University, Khorramabad, Iran. They were the students of four intact classes available to the researcher. Their age ranged from 18 to 24. All of them were female students.

Instruments

Two types of instruments were used in this study. The first instrument was a70-item CBT consisting of two sub-tests of receptive collocational behavior (hereafter RCBT) and productive collocation behavior (hereafter, PCBT).the items of CBT were selected from the British National Corpus (BNC).The BNC was built between 1991 and 1994 with a 100 million-word collection of samples of written and spoken

language from a wide range of sources, was designed to represent a wide cross-section of British English from the later part of the 20th century. The spoken component of the BNC constitutes approximately 10 percent (10 million words) and the written component 90 percent (90 million words) of the total data. There are nine written domains in the corpus: applied science, arts, belief and thought, commerce and finance, imaginative, leisure, natural and pure science, social science, and world affairs.

The second instrument was a validated Criterion Collocation Test (CCT) developed by Chen (2008) to assess the English collocation competence of college students in Taiwan. The CCT is a 50-item multiple choice test including verb, adjective, and proposition items. The validity of this test was presupposed as it was already checked by the developer. This test was run as a criterion measure against which the concurrent validity of the CBT was established. In this study, the reliability estimate of CCT was reported to be 0.81.

Procedure for the Operation of Different Levels of Inference

For this study, different levels of inference are presented with their own warrants, assumptions and backings. Both theoretical and empirical evidence are presented for different levels of inferences.

Observation Inference

The observation inference assesses the consistency of scoring methods with measurement processes. This inference is based on the warrant that the practice of reporting CBT scores as an overall indicator of students' collocational ability is fair and accurate. Here claims are supported by theoretical evidence represented in the formative stage of the interpretive argument. This warrant assumes that collocational language ability is better explained if the actual items in the test correspond to test specification. These warrants and assumptions are, in turn, supported by the backing illuminating that detailed procedure of test development as well as the results of expert review will reveal that test specification matches the items actualized in the finalized test.

Generalization Inference

The generalization inference rests on the warrant that expected scores are comparable across the items on the test. This inference is based on the warrants 1 & 2 that:

1. The generalizability of the CBT observed scores of the whole test as well as its two sub-tests suggest that CBT test scores can be used to generalize to a larger test population.
2. The desired reliability is attributable to the careful and systematic corpus-driven design and perhaps, to the construction of the test items.

These proposed warrants are formulated in the following assumptions:

1. The CBT distinguishes among test takers with acceptable difficulty and discrimination indexes.
2. CBT shows desired estimates of reliability for the whole test as well as the sub-tests.

Three types of backing support the assumptions proposed for the generalization inference. These empirical sources of evidence which are claimed to be at the summative stage of the interpretive argument are presented as follows:

1. Using Ebel's criteria, estimates from individual item analysis will reveal moderate estimates of difficulty and discrimination indices.
2. The results of *Kuder-Richardson (KR-21)* statistical analysis show that reliability coefficient of the whole SPT as well as its sub-tests is satisfactory, going beyond .8-a conventional yardstick against which reliability is measured (Jaen, 2007, p. 140).
3. Results of Pearson correlation shows a good estimate of reliability for CBT across forms

Extrapolation Inference

This inference is based on the warrant that the construct of collocational ability as assessed by the CBT accounts for the performance on an outside criteria. This inference warrants that in an EFL context CBT

provides an effective indicator of collocational language ability, assuming that CBT which is supposed to assess learners' collocational ability is a representative sample from the domain of interest. These assumptions and warrants are supported by the backing that Pearson Product coefficient will reveal a significant correlation between the scores on CBT and other measures of the same language ability.

Explanation/Implication Inference

Explanation inference can be evaluated from a theory-based perspective since it needs evidence to show the extent to which the construct and performance are relevant to a specific discipline. The warrant formulated in this inference shows that expected scores are attributed to not only the total test but to receptive and productive sub-tests of collocational behavior, claiming that both sub-tests of collocational behavior have a positive relationship with each other and with the total collocation test. And this is supported by Pearson–product correlation which reveals that both sub-tests of collocational behavior have a positive relationship with each other and with the total collocation test.

Table 2. *Summary of Warrants, Assumptions, and Backings of CBT (adapted from Chapelle, Enright & Jamieson, 2007b, p. 24)*

Inference	**Warrant Licensing the Inference**	**Assumptions Underlying Inferences**	***Backing***
Observation	The practice of reporting CBT scores as an overall indicator of students' collocational ability is fair and accurate.	Collocational language ability is better explained If the actual items in the test correspond to test specification.	Detailed procedure of test development as well as the results of expert review will reveal that test specification matches the items actualized in finalized test.

Generalization	1.The generalizability of the CBT observed scores of the whole test as well as its two sub-tests suggest that CBT test scores can be used to generalize to a larger test population. 2. The desired reliability is attributable to the careful and systematic corpus-driven design and perhaps, to the construction of the test items.	1. The CBT distinguishes among test takers with acceptable difficulty and discrimination indexes. 2. CBT shows desired estimates of reliability for the whole test as well as the sub-tests.	1. Using Ebel's criteria, estimates from individual item analysis will reveal moderate to high estimates of difficulty and discrimination indices. 2. The results of *Kuder-Richardson (KR-21)* statistical analysis show that reliability coefficient of the whole SPT as well as its sub-tests is satisfactory, going beyond .8-a conventional yardstick against which reliability is measured 3. Results of Pearson correlation shows a moderate reliability for CBT across forms
Extrapolation	In an EFL context CBT provides an effective indicator of collocational language ability	The CBT which assesses learners' collocational ability is a representative sample from the domain of interest.	Pearson Product coefficient will reveal a significant correlation between the scores on CBT and other measures of the same language ability.
Explanation	Expected scores are attributed to both receptive and productive Collocational behavior	Both sub-tests of collocational behavior have a positive relationship with each other and with total collocation test	Pearson –product correlation reveals that Both sub-tests of collocational behavior have a positive relationship with each other and with total collocation test

Results and Discussion

Research Question 1: Observation Inference

Are the procedures for CBT development detailed and based on test tasks presented in test specification?

To answer this research question, two theoretical levels of evidence were investigated. They are detailed procedure of test development, including the process of test specification (see table 3 below) and test construction as well as the analysis of expert review on these detailed procedures, the results of which are analyzed and discussed below.

Detailed procedure of item selection and test development (Theoretical evidence1).

Through the course of test development, I selected the word combinations (collocations) from the BNC corpus so that I could be able to include the target language use domain of collocational ability in the content of the test. As such, a 70-item test of collocational behavior (subsisting of a receptive and a productive subtest) was designed. For the receptive tasks, multiple-choice format and matching items were used. In this case, students were presented with the definitions of the concepts taken from *Collins Cobuild English Dictionary (2006)*. An example of an item for multiple-choice receptive tasks is presented below.

(ex.1) I have always enjoyed eating a substantial..... in a northern restaurant

a. food b. meal c. cake d. none of these

For the assessment of candidates' *productive* knowledge of collocational behavior, filling-in and translation tasks were included. Students were asked to complete a definition of the concept expressed by the intended collocations. This was, for example, the case in the following item of the productive CBT.

(ex.2) A daunting ……is the one in which people feel nervous and less confident to do it.

For translation task, however, some incomplete English statements (with their base nouns left out) were provided. The complete Persian translations of the statements were also presented. The base nouns in Persian were underlined and the subjects were required to fill in the blanks with appropriate English equivalents for the underlined base nouns in Persian.

Table 3. *Content Specification of CBT*

Components	No. of items	Points	Time (minutes)
Part I: receptive			
Section A: classifying	10	10	10
Section B: multiple-choice	20	20	20
Section C: matching	10	10	10
Part II: productive			
Section A: gap-filling (without contextualization)	10	10	10
Section B:gap-filling (contextualized)	10	10	10
Section C: translation	10	10	10
Total	70	70	70

Expert review of the match between test specification and actual test: (Theoretical Evidence 2)

Results of evaluation ratings. Three applied linguistics experts provided their feedback and suggestions on the preliminary test specification fit. Table 4 presents the results of the experts' ratings for each evaluation question. As shown in the Table, the experts generally agreed that the content domains and learning goals listed in the preliminary blueprint represent the target domains actualized in the final test. It also appeared that the tasks included are adequate and appropriate to assess students' collocational language ability. Only in two cases, one of the experts expressed dissatisfaction (items 4 & 8 in Table 4).

Table 4. *Expert Review of the Match Between Test Specification and Actual Test*

		Ratings Made by Experts			
Item	Evaluation Question	Strongly Agree	Agree	Disagree	Strongly Disagree
1	The receptive tasks represent the construct of collocational language ability.	X	XX		
2	The productive tasks included represent the construct of collocational language ability.	X	XX		
3	The words selected from the BNC are appropriate for developing items to assess students' receptive collocational language ability.	X	XX		
4	The words selected from the BNC are appropriate for developing items to assess students' productive collocational language ability.	X	X	X	
5	The domain of selected collocations are similar to those of target language use domain.	X	XX		
6	The domain of collocations are clearly described		XXX		
7	The corpus selected for test construction is appropriate for this goal.		XXX		
8	The items in the test specification fit those represented in the actual test.	X	X	X	

Research Question 2: Generalization Inference

To what extent items on the test distinguish among test takers with acceptable discrimination and reliability.

To answer this research question, the study made use of three empirical backings to systematically support the generalization inference: item analysis, KR-21 Reliability, and parallel-form reliability. Each will be analyzed and discussed below.

Item analysis (Empirical evidence 1).

As shown in Table 5, after an analysis of item difficulty, 3 items (all of them belonging to the productive collocation behavior test (PCBT) obtained p-values of .0, since they elicited incorrect answers from all the participants. As expected, the discrimination index showed that these highly difficult items were non-discriminating among candidates, and so they would need to be discarded in subsequent studies.

Of the 70 items included in the CBT, 12 (16%) were characterized as very difficult, 21 (30%) as difficult, 28 items (40%) offered a desirable level of difficulty, 8 items (9%) were easy and finally 1 item (1%) was classified as very easy items (Table 5). As such, about 60% of the items yielded a desirable level of difficulty, thus supporting the assumption that CBT items reveal a moderate level of difficulty and discrimination. It can be concluded that minimal construct irrelevant variance is introduced to the test construct.

Table 5**.** *Analysis of Difficulty of Individual Items According to Ebel's Criteria.*

Non-discriminating items (p value = .0)	CBT	RCBT	PCBT
	3 (4%)	0 (0%)	3 (4%)
Very difficult items (p-values from .01 to .14)	12 (17%)	0 (0%)	12 (17%)
Difficult items (p-values from .15to .39)	18 (25%)	8 (11%)	10 (14%)
Desirable items (p-values from .40 to .70)	28 (40%)	24 (34%)	4 (5%)
Easy items (p-values from .71 to .85)	8 (11%)	7 (10%)	1(1%)
Very easy items (p-values from .86 to 1)	1 (1%)	1(1%)	0 (0%)

Reliability is claimed to correspond to generalization inference. Two empirical sources of evidence provided backing for this inference.

KR-21 & parallel-form reliability (Empirical evidence 2& 3).

Two methods of *Kuder-Richardson (KR-21)* and *parallel-form reliability* were applied to establish the desired reliability of CBT. The calculated reliability estimate for CBT was .84. The reliability for

receptive and productive sub-tests were estimated to be .82 and .61 respectively. These results show that reliability coefficient of the whole CBT is satisfactory, going beyond .8-a conventional yardstick against which reliability is measured (Jaen, 2007, p.140). It can be argued that careful and systematic corpus-driven design of the test may be the reason for this high estimate of reliability. As for receptive CBT it also holds true since again the reliability coefficient goes beyond the specified yardstick. However, for productive CBT the reliability coefficient is reported to be less satisfactory. This can be partially attributed to the smaller number of items (30) as compared with the receptive test, or it may be due to the fact that productive items are more difficult and demanding than their receptive counterparts, thus leading to less variability among the scores and consequently to lower reliability. Overall, the whole test was reliable and enough to support the assumption that CBT shows a desired estimate of reliability.

To establish the reliability of the test across forms (A and B), parallel-form method was used. The correlation between CBT A and CBT B was found to be 0.39. This value is significant at the level of 0.03 showing a meaningful correlational coefficient between the two versions of CBT, hence rendering the test as reliable, and adding another empirical backing to the network of generalization inference.

Table 6. *Correlation Between CBT A and CBT B*

		CBT A	CBT B
CBTA	Pearson correlation	1	0394٠(**)
Sig. (2- tailed)		.	٠.031
N		60	60
CBT B	Pearson correlation	0394(**)	1
	Sig. (2- tailed)	0.031	.
	N	۶۰	60

**Correlation is significant at the 0.05 level (2-tailed).

Research Question 3: Extrapolation Inference

Do test-takers' performance on the collocation test correlate positively to performance on a concurrent measure of English ability as expected?

This research question addressed the argument that the CBT can be extrapolated to non-test behavior or to a criterion measure and was examined through the use of Pearson-Product correlation coefficient.

Correlation between CBT and a concurrent measure (Empirical evidence 4).

In extrapolation inference it is hypothesized that if the newly developed test (CBT) is a valid measure of collocational behavior, it will correlate significantly with the outside criterion measuring the same construct. For this purpose, the scores on CBT and a criterion measure known as CCT were correlated to each other (see Table 7). The result showed a significant correlation coefficient of 0.29, $p< 0.05$. This finding indicates a desired correlation and is documented as an empirical evidence to support the extrapolation inference in the interpretive argument, being consistent with Bachman's (1990) interpretation that some correlations, if moderately high, can be cited as evidence that the new test measures approximately the same general area of behavior as other tests designed by the same name as the new test.

Table 7. *Correlation Between CBT and CCT*

			CBT	CCT
CBT	Pearson	correlation	1	.289 (**)
Sig. (2- tailed)			.	.093
N			60	60

** Correlation is significant at the 0.05 level (2-tailed).

Research Question 4: Explanation Inference

Is there a strong correlation between scores on the collocation test and scores on the receptive and productive sub-test of collocation?

Internal test structure refers to the interrelationships, often expressed as correlations or covariance, between performance on different parts of a single test form. Empirical evidence or backing for the explanation inference addressed in the above question came from the internal consistency of CBT and its sub-tests, often actualized through the use of correlations and inter correlations. When examining the internal

structure of an assessment, the extent to which the individual items and the assessment, itself, measure the intended construct(s) is of primary interest. As indicated by the *Standards for Educational and Psychological Testing* (1999), evidence based on the internal structure of the assessment indicates "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 13).

Internal consistency of CBT and its sub-tests (Empirical evidence 5). Referring to the information reported in Table 8, the correlation coefficient between the two sub-tests (receptive CBT and productive CBT) is .45. Between the productive CBT and the total CBT this value is .74, which is a high estimate. Interestingly, the coefficient value between RSP and total SPT is .92 which is the strongest, as compared to the other two correlations. Thus, it can be argued that CBT portrays lower correlation to the Productive CBT than to the Receptive one. One source of explanation is that difficult nature of productive items may lower the correlation value of this sub-test with the receptive or total test items. This finding is in line with Jaen's (2007) study wherein he concluded that learners have more problems with producing collocations than with recognizing them. All in all, by relying on these values, we can safely claim that the test supports the explanation inference of the interpretive argument.

Table 8 *Internal Consistency of CBT and the Sub-Tests*

		CBT	RCBT	PCBT
CBT	Pearson correlation	1	.923(**)	.740(**)
	sig.(2-tailed)	.	.000	.000
	N	60	60	60
RCBT	Pearson correlation	.923(**)	1	0.454 (**)
	sig.(2-tailed)	.000	.	.000
	N	60	60	60
PCBT	Pearson correlation	.740(**)	0.454 (**)	1
	sig.(2-tailed)	.000	.000	.
	N	60	60	60

Note. ** Correlation is significant at the level of 0.05 (2-tailed).

Conclusions and Implications

Relying on multiple sources of evidence or backing, we can conclude that not only the newly developed CBT can adequately measure students' level of collocational ability and can provide useful information for formative assessment to understand students' current standing on collocational knowledge but it stood all the tests of argument based validity. Simply put it, all the assumptions of CBT were supported.

As such, the present study has the potential to make contributions to the study of collocations and validation of language assessment. The network of inferences together with their corresponding backing proposed in this study can lead to a better-informed test development in an argument-based validity approach. Findings from the present and similar studies (e.g., Abdul Kadir, 2008; Chapelle, Enright and Jamieson 2004, 2008, 2010; Chung, 2014; Jun, 2014; Liu, 2013; Johnson & Riazi, 2013; Xi, 2008) which follow an argument-based approach to validity would be enlightening for test developers, since they would not be bothered by the limitations of unidimensinality and unsystematicity as attributed to traditional approaches to validity. Proposing assumptions and providing backing in the form of arguments gives test developers and other stake holders a better picture of validity generally and a more robust interpretation of test use specifically.

This study also paves the way for further investigations to be made in this area. One area of research area, for example would be an assessment on collocational behavior which is based on multiple sources of evidence informing both test interpretation and test use, an under researched area in validity studies. Also, the triangulation of different sources of evidence as well as getting insight and perspectives from different stakeholders such as academic experts, teachers, test-takers, and other test users should be regarded in later investigations of test development in argument-based approaches to validity. Finally, embarking on a mixed method approach to gain a better panorama of some of the facets of validity is urgent. Hope this study contributes to the betterment of test development in general and to the collocation tests in particular.

References

Abdul Kadir, K. (2008). *Framing a validity argument for test use and impact: Malaysian public service experience.* Unpublished Doctoral dissertation, University of Illinois at Urbana-Champaign.

Almela, M. (2007). Words as "lexical units" in learning teaching vocabulary. *International Journal of English Studies, 7*(2), 21-40.

Aryadoust, V. (2011).validity arguments of the speaking and listening modules of international English language testing system: A synthesis of existing research. *Asian ESP Journal.7(2), 28-54.*

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

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1-34.

Barfield, A. (2003). *Collocation recognition and production: Research insights*. Tokyo: Chuo University.

Bonk, W.J. (2001). Testing ESL learners' knowledge of collocations. In T. Hudson &J. D. Brown (Eds.), *A focus on language test development: expanding the language proficiency construct across a variety of tests. (Technical eport#21).* Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center, (pp. 113-142).

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10(2),* 157-187.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2004). *Issues in developing a TOEFL validity argument: Paper presented at the 26th Annual Language Testing Research Colloquium*, Temecula, CA.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2007a). *(Where) is the construct in an Interpretive Argument? Paper presented at the 29th Annual Language Testing Research Colloquium*, Barcelona, Spain.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2007b). *From validation research to a validity argument. Paper presented at the 4th European Association for Language Testing and Assessment Conference,* Sitges, Spain.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language.* London: Routledge.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3- 13.

Chapelle, C. A., & Read, J. (2001). A framework for vocabulary assessment. *Language Testing, 18(1),* 1-32.

Cronbach, L. J. (.1971). Validity. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-597). Washington, DC: American Council on Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Oxford, UK: Routledge.

House, E. R. (1980). *Evaluating with validity. Beverly Hills*, CA: SAGE.

Jaen, M.M. (2007). A corpus-driven design of a test for assessing the ESL collocational competence of university students. *International journal of English Studies, 7*(2), 127- 147.

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21(1),* 31-41.

Kane, M. (2006). Validation. In R. Brennen (Ed.), *Educational measurement* (4th ed.), (pp. 17-64). Westport, CT: Greenwood.

Kane, M. T. (2006a). Content-related validity evidence. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–154). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2006b). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17- 64). American Council on Education/Praeger.

Kane, M.T.(2011).Validating score interpretations and uses. *Language Testing, 29*(1) 3– 17

Keshavarz, M. H. & Salimi, H. (2007). Collocational competence and cloze test performance: *A study of Iranian EFL learners*, *17*(1), 81-92.

Le, H.T. (2011). *Developing a validity argument for the English placement Fall 2010 Listening test at Iowa State University,* Unpublished Doctoral Dissertation, Iowa State University.

McNamara, T. F., &Roever, C. (2006). *Language testing: The social dimension.* Maiden, MA: Blackwell Publishing.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13 103). New York: American Council on Education and Macmillan

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.

Mochizuki, M. (2002). Exploration of two aspects of vocabulary knowledge: Paradigmatic and collocational. *Annual Review of English Language Education in Japan, 13*, 121- 129.

Park, J. (2012). *Developing and Validating an Instrument to Measure College Students' Inferential Reasoning in Statistics*: An Argument-Based Approach to Validation. Unpublished Doctoral dissertation, University of Minnesota.

Pawley, A., &Syder, F. H. (1983). Two puzzles for linguistic theory: Native like selection and native like fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). New York, NY: Longman.

Qian, D. &Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing, 21*(1), 28-52.

Johnson, R.C &Riazi, M.(2013). Assessing the assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context. *Papers in Language Testing and Assessment, 2(1)*, 31-58.

Toulmin, S. E. (2003). *The uses of argument (updated edition).* Cambridge, UK: Cambridge 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*. Unpublished doctoral dissertation, Iowa State University.

Widdowson, H. G., (2007). *Discourse analysis*. Oxford: Oxford University Press.