Psychological Test and Assessment Modeling, Volume 56, 2014 (1), 60-82

Establishing the construct validity of conversational C-Tests using a multidimensional Rasch model

Purya Baghaei[1] *& Rüdiger Grotjahn*[2]

Abstract

C-Test is a variation of cloze test where the second half of every second word is deleted. The number of words correctly reconstructed by the test taker is considered to be a measure of general language proficiency. In this pilot study the componential structure of an English C-Test consisting of two spoken-discourse passages and two written-discourse passages is investigated with the help of both unidimensional and multidimensional Rasch model. In a sample of 99 fairly advanced Iranian students of English the data fitted better the multidimensional partial credit model as defined in multidimensional random coefficients multinomial logit model (Adams, Wilson, & Wang, 1997) than Masters' (1982) unidimensional partial credit model. This indicates that spoken-discourse and written-discourse C-Test passages form distinct dimensions. We argue that spoken-discourse C-Test texts may tap better into students' listening/speaking skills than C-Test based solely on written discourse texts and that therefore C-Tests consisting of both conversational and written-discourse passages can more adequately operationalize the construct of general language proficiency than C-Tests containing only written discourse passages. Considering the small sample size of the study the findings should be interpreted cautiously.

Key words: multidimensional random coefficients multinomial logit model, C-Test, MIRT, structure of general language proficiency

[1] *Correspondence concerning this article should be addressed to:* Dr. Purya Baghaei, Islamic Azad University, Faculty of Foreign Languages, Ostad Yusofi St., 91886 Mashhad, Iran; email: pbaghaei@mshdiau.ac.ir

[2] Ruhr-Universität Bochum, Seminar für Sprachlehrforschung, 44780 Bochum, Germany

Introduction

The C-Test is a variation of the cloze test. In its original form, a cloze test consists of a single long passage in which, after a short unmutilated lead-in, every *n*th word is deleted, *n* being usually a number between 5 and 10. Test takers have to insert the missing words (cf. Grotjahn, 2013). C-Tests, in contrast, always consist of several short passages (most often four to six) in which the second half of every second word is deleted. Usually there are 20–25 mutilated words in each passage. Every word that is correctly reconstructed by the test takers is scored one and otherwise zero. As a rule, C-Tests which have been carefully designed and pretested are highly reliable. Especially in the case of longer C-Tests such as the onDaF (see https://www.ondaf.de/gast/ondaf/info/home.jsp), reliability coefficients often exceed 0.9, and even in the case of shorter C-Tests they often exceed 0.85 (see Eckes, in press; Eckes & Grotjahn, 2006; Grotjahn, in press). C-Tests are considered to be tests of general proficiency and are widely used in language testing (see Grotjahn, in press and http://www.c-test.de/).

Since in a cloze test there is only one passage the chances that some examinees will be familiar with the content of the passage and thus favored are high compared to a C-Test with 4–6 passages. In addition, because of local dependence of items in cloze tests the application of KR-20 and Cronbach's Alpha formulas to estimate reliability is problematic because these formulas assume local independence of items. Furthermore, the application of the Rasch model and other IRT (item response theory) models is also problematic with cloze because of the same reason. In C-Tests, however, each passage is normally considered a super-item and entered into the analysis as a single independent item. This makes the application of item response models assuming local item independence possible in C-Test analysis. Recently, as an alternative, Harsch and Hartig (2010) have used the Rasch testlet model (Wang & Wilson, 2005) and Eckes and Baghaei (accepted, in press) have used testlet response theory (Bradlow, Wainer, & Wang, 1999) to take account of local item dependence in C-Tests. These models are analogues of the standard Rasch and IRT models with an extra random effect parameter for testlets.

Rationale for the study

As already mentioned C-Tests are considered to be tests of general language proficiency and test takers' scores on C-Tests are interpreted as indicators of their overall ability in a (foreign) language. However, such an interpretation may only be partially warranted since C-Tests do not engage students in oral/aural skills (cf. Shohamy, 1982, p. 162; and also Eckes & Grotjahn, 2006, p. 297). One could therefore argue that C-Test scores might be better indicators of reading/writing competence than of listening/speaking ability.

In line with this argument, Alderson (2002) explicitly claimed that the C-Test is not an adequate measure of general language proficiency because it does not tap into oral-aural skills. Concurrent validation studies corroborate this claim to a certain degree. For example, Chapelle and Abraham (1990) reported a correlation coefficient of 0.47 (corrected for attenuation) between an English C-Test and the listening section of the Iowa State

English Placement Test. Dörnyei and Katona (1992) reported correlation coefficients of 0.33 and 0.51 between an English C-Test and both the listening sections of a university department test and the Test of English for International Communication (TOEIC). The authors also observed a correlation coefficient of 0.43 between the C-Test and an oral interview. Grotjahn (1992) found a correlation coefficient of 0.24 between a French C-Test and students' self-ratings of their speaking ability. Coleman (1994), however, reported a much higher correlation coefficient of 0.76 for listening comprehension and a moderate coefficient of 0.47 for speaking (Cambridge A-level examinations). Arras, Eckes and Grotjahn (2002) found a correlation coefficient of 0.64 between a German C-Test and both the speaking and the listening subtests of TestDaF (Test Deutsch als Fremdsprache; Test of German as a Foreign Language), whereas Eckes (in press) found correlation coefficients of 0.62 and 0.48 between the onDaF C-Test system and the listening and speaking subtests of TestDaF. Daller and Phelan (2006) reported a correlation coefficient of 0.45 for an English C-Test and the TOEIC listening section. The results of correlational studies of C-Tests and tests of speaking and listening are thus inconsistent, but low correlations are more frequently reported than high correlations and the correlation coefficients reported for speaking are lower than those for listening (for more information on correlational studies see Eckes & Grotjahn, 2006).

C-Test texts are normally taken from written discourse. This is also true for the studies cited above. A notable exception is Daller and Grotjahn (1999) who tried to measure two separate dimensions, namely "academic language proficiency" (ALP) and "everyday language proficiency" (ELP). For measuring ALP they used texts from university textbooks and for measuring ELP they used texts from newspapers about everyday topics. Using both factor analysis and the Classical Latent Additive Test Model (cf. Moosbrugger & Müller, 1982) Daller and Grotjahn (1999) could establish that ALP and ELP constitute two different dimensions (cf. also Baghaei & Grotjahn, in press).

In the present study we tried to find out if it is possible to distinguish two distinct dimensions of reading/writing skills and listening/speaking skills in C-Tests composed of both spoken discourse and written discourse passages. If research shows that this is indeed possible, then a limitation of C-Tests as an indicator of general language proficiency could be overcome by constructing C-Test batteries which contain both written and spoken language texts. In addition, when the aim is to predict more precisely the oral/aural competence of examinees one could also consider using C-Tests which consist only of spoken discourse passages.

Moreover, the inclusion of spoken discourse texts increases content validity and by implication construct validity and generalizability of score interpretation (Baghaei, Monshi Toussi, & Boori, 2009). Even in direct tests of language skills the advice for test developers is to include as many different tasks and texts with different styles, topics and purposes. In writing tests for example, examinees should be required to write at least two or three different pieces (e.g., a letter and a post card), so that the test user can make sound interpretations about the test taker's writing ability.

The research mentioned so far is mainly correlational. We will now briefly comment on some other pertinent research.

Carrell (1984), for example, using an experimental design, discovered that rhetorical organization of texts affected reading comprehension and text recall of intermediate non-native readers of English. She noticed that students performed differently on texts which had different structures such as problem/solution vs. comparison. This is in line with numerous other studies on first and second language text processing (for more information on the effect of text type on reading comprehension cf., e.g., Grabe, 2009).

Shohamy and Inbar (1991) in testing listening comprehension used three different text types and two different topics. The text types varied in the degree of oral features they contained: a) conversation, b) lecture, and c) news broadcast. Then identical listening comprehension questions were constructed for the three genres. Although the three text types on the same topic contained the same information and also the listening comprehension questions for all the texts were identical, students performed differently on the text types.

Referring to this and other research, Baghaei (2008) analyzed a C-Test battery composed of four passages with four different rhetorical organizations: description, causation, comparison, and problem/solution. The results of the study showed that rhetorical organization affected the skills needed for solving C-Tests. He concluded that what C-Tests measure depends to a considerable extent on the rhetorical structure of the texts used.

Equally relevant are the numerous analyses of C-Tests based on classical test theory and item response theory (IRT), which show that there are always some texts in C-Test batteries which should be discarded because of lack of fit, low text-total correlations or low factor loadings. This is evidence of the effect of the nature of text on the kind of abilities which are triggered by C-Tests. This has always been a nuisance for C-Test users. However, a closer look at this phenomenon can lead to a more informed construction and application of C-Tests.

Finally we would like to mention Sigott's (2004) notion of the fluid construct phenomenon in C-Tests. What he basically means by this notion is the fact that the C-Test construct is not stable and changes as a result of text difficulty, test taker ability and other characteristics of the test taker. His main focus was on the potential of C-Tests for involving test takers in high-level and low-level skills. He found that high-ability students managed to solve many C-Test items even when these were decontextualized, whereas low-ability test takers needed much more context for solving the items. He therefore argued that for high-ability test takers the C-Test is a test of lower-order skills, whereas for low-ability test takers it is a test of higher-order skills. This means that conclusions as to what a C-Test measures should be qualified at least with regard to the ability of the test takers and the difficulty of the texts.

Altogether, the findings of statistical analyses of C-Test data including correlational, factor analytic and IRT research as well as (experimental) research in text linguistics and schema theory suggest that in C-Test processing the features of texts do play a role in what the test measures and thus affect construct validity. This being the case, we can hypothesize that constructing C-Tests on the basis of spoken discourse passages may lead to C-Tests which tap into listening/speaking ability to a larger extent than tests consisting exclusively of written discourse passages.

In this pilot study we aim to demonstrate the application of multidimensional Rasch modeling to examine the validity and dimensionality of language tests in general and C-Tests in particular. We specifically focus on the dimensionality of C-Tests composed of texts with different genres and stylistic features and suggest that information on the effect of genre and stylistic features should be taken into account when reporting results from C-Tests. This is in line with Messick (1989) who, referring to the structural aspect of test validity, states that the scoring profile of test data should be informed by the componential structure of tests. As the sample size of the present study is fairly small, the results should be interpreted with caution.

Before describing our own study, in which C-Tests consisting of both oral and written discourse passages are analysed with the help of unidimensional and multidimensional IRT, we first present a brief introduction to multidimensional IRT modeling.

Multidimensional IRT models

Basic characteristics

One of the most fundamental features of traditional IRT models is the unidimensionality assumption. That is, all the items in an instrument should measure one single trait or dimension. Traditionally, measurement in humanities and physical sciences has been unidimensional. In other words, the aim is to measure one single variable at a time. We normally do not conflate measures. This is essential for measurement; otherwise comparison of respondents' ability or trait measures is not possible and any further generalizations and inferences based on their measures are misleading (cf. Hattie, 1985 for further information).

Establishing unidimensionality (see Kubinger, 2005 for an overview of Rasch model checks), however, is very difficult or even sometimes impossible in practical testing situations. The unidimensionality assumption is thus violated very easily. Moreover, in many educational measurement contexts we deliberately construct multidimensional tests to measure students' abilities on several dimensions. Multidimensional IRT models (MIRT) account for multiple dimensions in a dataset and estimate students' abilities separately on the dimensions involved. Fitting a unidimensional model to a multidimensional test results in loss of information and disappearance of subscales. As a consequence we cannot investigate possible relationships among dimensions (Adams, Wilson, & Wang, 1997; Brandt, 2012; Höhler, Hartig, & Goldhammer, 2010).

An alternative to MIRT, which can be used to study multidimensional tests, is the consecutive approach suggested by Davey and Hirsch (1991, cited in Adams et al., 1997). In this approach each dimension of the test is analysed separately with a unidimensional model. This approach has certain advantages over a joint unidimensional analysis where all items are considered to belong to a single dimension: it recognizes the multidimensionality of the instrument and provides a measure on each dimension and thus also allows us to investigate the relationships among the dimensions. One drawback to this approach is the measurement error which is associated with the estimates of items and

persons on each dimension. Since the subsets which form the dimensions often consist of only a few items, measurement error will be large. Subscales of at least 20 items are needed to have an acceptable error of measurement (Adams et al., 1997).

Another drawback of the consecutive approach is its failure to use all available data (Adams et al., 1997). In other words the approach uses only the portion of the data which is related to a dimension and ignores the rest. A combined analysis with a multidimensional model results in more stable and accurate item and person parameter estimates because all available data are used.

We can also look at multidimensional IRT models through the lens of classical factor analysis (FA). Reckase (1997) considers MIRT as a special case of FA because both try to detect hypothetical scales or factors on the basis of a matrix of observed scores. However, the focus of the two approaches is quite different; while FA attempts to reduce the data to a minimum number of underlying factors, MIRT tries to parameterize items and persons on a common scale so that one can predict the chances of success of a person with a known ability parameter on an item with a known difficulty parameter. That is, in MIRT models we want to model the interaction between persons and items in order to understand the characteristics of persons and items and the nature of their interaction (Reckase, 1997). The more dimensions we extract from the data the more precise our understanding of the nature of the interaction will be. Therefore, one of the basic distinctions between MIRT models and FA is that MIRT models do not focus on reducing the data to a minimum number of underlying factors. MIRT accounts for profiles of proficiency rather than overall proficiency and, as was mentioned above, items can measure one or more latent dimensions.

Adams et al. (1997) summarize the advantages of analyses based on multidimensional models as follows:

1. They take care of the intended structure of the test in terms of the number of subscales.
2. They provide estimates of the relationships among the dimensions.
3. They make use of the relationships among the dimensions to produce more accurate item and person estimates.
4. They are single rather than multistep analyses.
5. They provide more accurate estimates than the consecutive approach.
6. Unlike the consecutive approach they can be applied to tests that contain items which load on more than one dimension.

Multidimensional models for confirmatory analyses

MIRT models can be used in a confirmatory mode to investigate the construct validity, and in particular the cognitive validity, of tests with multiple components or subscales (Baghaei, 2013; Embretson, 1980, 1983; Field 2013; Janssen & De Boeck, 1999; Santelices & Caspary, 2009; Wilson & Moore, 2012). The components or subscales are considered to be different cognitive processes or abilities which are involved in solving the

items. Traditionally, factor analytic approaches have been used to study the componential structure of tests. MIRT methods, on the other hand, model the interaction between respondents and tasks to identify components. In factor analytic techniques the components are identified on the basis of correlations between subscales. These correlations, however, represent many things such as:

... knowledge prerequisites, educational communalities, and genetic communalities, as well as common underlying theoretical components ... [the factors] represent influences that cannot be separated in a given set of variables, but this does not necessarily imply that they represent elementary theoretical mechanisms. Factor analysis is unable to separate one or more unique "components" in a task from error variance (Embretson, 1983, p. 180).

Multidimensional Rasch and IRT models have some advantages over factor analysis for dimensionality assessment. In IRT models information from examinee response patterns is analysed while in factor analysis information from correlation matrices is analysed which is more limited (Lane & Stone, 2006). Furthermore, nonlinear models such as IRT may better reflect the relationship between item performance and the latent ability (Hattie, 1985).

In modern assessment we are interested in more than summarizing test takers' abilities in a unidimensional single measure. We are interested in grasping what kind of knowledge, strategies and skills are used by test takers in responding to test items (Adams et al., 1997; Embretson, 1983; Hartig & Höhler, 2008, 2010; Koeppen, Hartig, Klieme, & Leutner, 2008). MIRT models can unfold these underlying strategies, skills and knowledge structures which is a great help in apprehending what goes on in the mind of the examinees when they tackle the test tasks, and therefore can be used as powerful statistical techniques for construct validation.

Investigation of the componential structure of tests is closely related to Embretson's notion of construct representation. Embretson (1983, 2007) states that construct representation as an aspect of internal validity relates to the processes, strategies and knowledge structures that examinees employ when they perform the test tasks, or in other words, to the meaning of test scores in terms of the processes, strategies and knowledge that the test triggers in the examinees. She also introduces the concept of nomothetic span as an external aspect of construct validity which deals with the relationship between test scores and other tests and external criteria.

Assessing construct representation by means of MIRT requires a well-designed theory-informed test in which the cognitive structures tapped by subscales have been hypothesized a priori (Embretson, 1985; Hartig & Höhler, 2010; Janssen & De Boeck, 1999; Wilson & Moore, 2011). MIRT models can then be applied to check the plausibility of the hypothesized structure in terms of model-data fit and information criteria. If a certain hypothesized test structure fits the MIRT model, or if a certain hypothesized test structure fits better than a competing alternative structure, one has adduced evidence for the appropriateness of the structure and the cognitive processes which were assumed to be triggered by each component.

Model selection criteria

Competing models are compared by calculating the likelihoods of their solutions. The greater the likelihood, or lower –2log-likelihood (deviance) the better the data fit the model. Therefore, we expect the deviance to be small and models with smaller deviances are selected (Janssen & De Boeck, 1999). Two nested models, where one model is a more constrained version of the other model, can be compared with the likelihood ratio test. The difference in the –2log-likelihoods should be distributed as χ^2 with degrees of freedom equal to the number of additional free parameters (DeMars, 2012). Therefore, nested models can be compared with a statistical significance test.

Information criteria are based on both the log-likelihood and the number of parameters estimated and thus take also the scientific criterion of parsimony into account. Akaike information criterion (AIC) suggested by Akaike (1974) is computed as:

$$AIC = -2 \log ML + 2p,$$

where *ML* is maximum likelihood and *p* is the number of estimated parameters in the model. The number of parameters is included in the model as a penalty term for overparameterization (Kang & Cohen, 2007). AIC is not asymptotically consistent as sample size is not used in its calculation.

Therefore, Bayesian information criterion (BIC) or Schwarz criteria was suggested by Schwarz (1978):

$$BIC = -2 \log ML + p (\log n),$$

where *n* is the sample size. BIC penalizes more for the number of parameters and hence favors models with fewer parameters compared to AIC. Models which have small AIC and BIC are selected. According to Lin and Dayton (1997) the results of the two statistics do not necessarily agree.

Method

Subjects

A sample of 99 Iranian students of English, comprising 23 males and 76 females aged between 21 and 33 ($M = 21.2$; $SD = 6.3$) participated in this study. All participants were undergraduate English majors at Islamic Azad University Mashhad Campus in Iran; almost all of them were in the third year of their studies.

Materials

For the purposes of this study a C-Test battery comprising four passages was constructed. Two passages were taken from dialogues available on ESL websites, one from www.1-language.com/ and another from www.focusenglish.com/. These spoken discourse passages are also referred to as conversational passages/texts. The two written discourse passages were taken from literary books. One is a short extract from George Orwell's Animal Farm and the other a paragraph from a book on Shakespeare's Sonnets by Ridden (1982). Each passage has 25 blanks totaling 100 in the entire C-Test (see Appendix).

Procedures

For data analysis we used the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM) (Adams et al., 1997). MRCMLM is a compensatory multidimensional model belonging to the Rasch model framework. Therefore, it has the measurement properties of these models such as specific objectivity (invariant comparisons) and sufficient statistics for parameter estimation. MRCMLM is a very flexible model which can accommodate, in addition to the original dichotomous model (Rasch, 1960/1980), a range of Rasch model extensions including the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982) and the facets model (Linacre, 1989). MRCMLM has also been used for the analysis of the PISA data (cf. Walter, 2005). MRCMLM is implemented in ConQuest (Wu, Adams, Wilson, & Haldane, 2007) using the Expected A Posteriori (EAP) estimation method (Bock & Aitkin, 1981), which utilizes correlations among dimensions to improve estimation accuracy.

MRCMLM can formally be expressed as follows:

$$p_{nij} = \frac{\exp\left(\mathbf{b}_{ij}'\boldsymbol{\theta}_n + \mathbf{a}_{ij}'\boldsymbol{\xi}\right)}{\Sigma_{u=1}^{k_i}\exp\left(\mathbf{b}_{iu}'\boldsymbol{\theta}_n + \mathbf{a}_{iu}'\boldsymbol{\xi}\right)}$$

where p_{nij} is the probability of a response in category j of item i for person n; person n's levels on the D latent variables are denoted as $\boldsymbol{\theta}_n' = (\theta_{n1},\ldots,\theta_{nD})$, k_i is the number of categories in item i, $\boldsymbol{\xi}$ is a vector of difficulty parameters, $\mathbf{b}_{ij}$ is a score vector given to category j of item i, $\mathbf{a}_{ij}$ is a design vector given to category j of item i that describes the linear relationship among the elements of $\boldsymbol{\xi}$ (Wang, Cheng, & Wilson, 2005, p. 14). The ConQuest programme (Wu et al., 2007) in which marginal maximum likelihood estimation is implemented was used to estimate model parameters.

Results

Check of unidimensionality versus multidimensionality

In our first analysis a unidimensional model was assumed to account for the data (Model 1). As the assumption of equal item thresholds cannot be assumed for C-Test passages, Master's (1982) partial credit model was used to analyse the data (cf. Eckes, 2006, 2011 for a comparison of various Rasch models for the analysis of C-Test data). In the second analysis, the data were analysed with the multidimensional random coefficients multinomial logit model (MRCMLM; Adams et al., 1997) as implemented in ConQuest (Wu et al., 2007). Here the two spoken discourse passages C1 and C2 were modeled to load on one latent dimension named Listening/Speaking and the two written discourse passages W1 and W2 were modeled to load on the second latent dimension named Reading/Writing (Model 2). The two models are depicted in Figures 1 and 2.

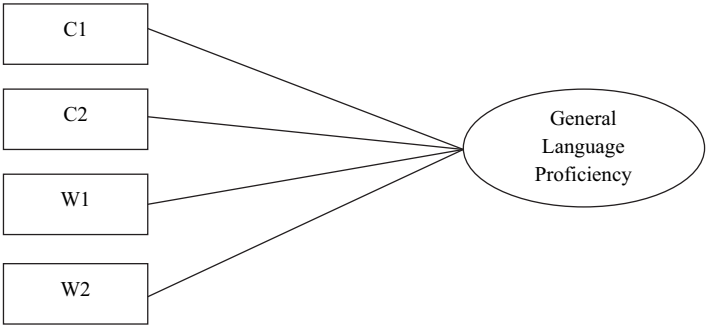

Figure 1:
Graphical representation of Model 1

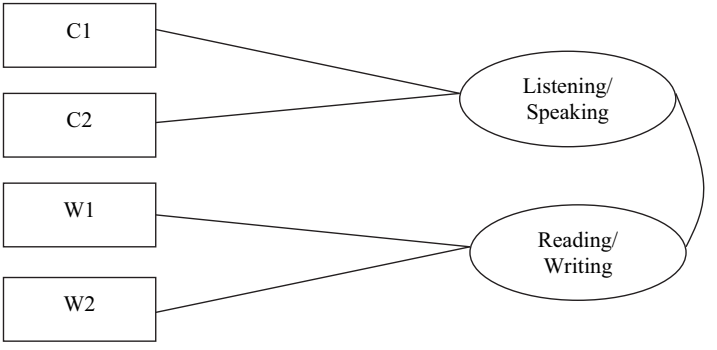

Figure 2:
Graphical representation of Model 2

Tables 1 and 2 below show the passage difficulty estimates, their standard errors and fit statistics in the two analyses. Infit and outfit mean square statistics (MNSQ) have an expected value of 1. Their corresponding *t* standardized values have an expected ideal value of 0; values between –2 and +2 are considered acceptable (Bond & Fox, 2007). As the fit values in Tables 1 and 2 clearly indicate, the two-dimensional model fits the data much better than the unidimensional model.

An overall model-data fit test was also used to compare the fit of the two models. This test is done by comparing the final deviance statistics of the two models. The model which has a smaller deviance has a better fit (Wu et al., 2007). The difference between the deviances of two models is assumed to be chi-square distributed with degrees of freedom corresponding to the difference between the numbers of parameters of the two models (Wu et al, 2007). The results of the overall model-data fit test are displayed in Table 3.

Table 1:
Passage statistics for the unidimensional analysis (Model 1)

C-Test Text	Estimate	SE	Infit		Outfit	
			MNSQ	*t*	MNSQ	*t*
C 1	–0.07	0.02	1.23	1.5	1.18	1.2
C2	–0.06	0.02	0.98	–0.1	0.99	0.0
W1	0.01	0.02	0.85	–1.0	0.85	–1.0
W2	0.13	0.04	0.98	–0.1	0.95	–0.3

Table 2:
Passage statistics for the two-dimensional analysis (Model 2)

C-Test Text	Estimate	SE	Infit		Outfit	
			MNSQ	*t*	MNSQ	*t*
C1	–0.004	0.02	1.04	0.3	0.99	0.0
C2	0.004	0.02	0.98	–0.1	0.97	–0.2
W1	–0.091	0.03	1.02	0.1	1.01	0.1
W2	0.091	0.03	1.05	0.4	1.03	0.3

Table 3:
Model fit statistics for the unidimensional and multidimensional model

Model	Deviance	Estimated Parameters	Correlation Between Dimensions	AIC	BIC
Unidimensional (Model 1)	2016.75	84	–	2184.75	2402.31
Two-Dimensional (Model 2)	2007.20	86	0.882	2179.20	2401.94

As Table 3 shows, the two-dimensional model has a smaller deviance statistic. The difference in the deviances is significant ($p < 0.01$; $df = 2$). This is an indication that the C-Test data fit the two-dimensional model better than the unidimensional model. This is corroborated by the information criteria AIC and BIC, which are both smaller for the two-dimensional model. Students' performance on a C-Test, which contains both written discourse passages and spoken discourse passages, can thus be modeled more efficiently if we consider the C-Test as two-dimensional and report two ability estimates for each student, i.e., one for each dimension. The claim that the two dimensions which are measured by the two text types are distinct is thus verified in this study notwithstanding that they are highly correlated ($r = 0.882$). In line with the theoretical arguments and empirical evidence adduced above we suggest that these results be interpreted that conversational C-Test texts tap into specific aspects of listening/speaking ability not accounted for by written discourse C-Test texts.

One of the important features of Rasch and IRT models is that person and item parameters are expressed on a common scale. This feature enables us to map person abilities and item threshold parameters to graphically display and compare the distribution of item parameters against person parameters. Such item-person maps help to assess whether the test is well-targeted for the ability distribution of test takers and whether all regions of the ability distribution are covered with items. Figure 3 displays the item-person map of the latent two-dimensional distribution of item thresholds and person abilities. Xs indicate person locations on the two dimensions; numbers on the right indicate the thresholds associated with each item. For example, '1.14' means threshold parameter 14 of Item 1. Threshold estimates can be read from the calibrated vertical line on the left. It can be seen that the item thresholds cover well the distribution of person abilities and that Dimension 2 (Reading/Writing) spreads the test takers more widely than Dimension 1 (Listening/Speaking). This also indicates that the two dimensions are distinct. Furthermore, the figure shows that the test is well-targeted for the sample, i.e., person abilities and item difficulties match. Kang and Cohen (2007) in a simulation study demonstrated that when test difficulty matches the distribution of examinee ability model selection is more accurate. This can be taken as a support that our model selection process by means of likelihood deviance test and information criteria is reliable.

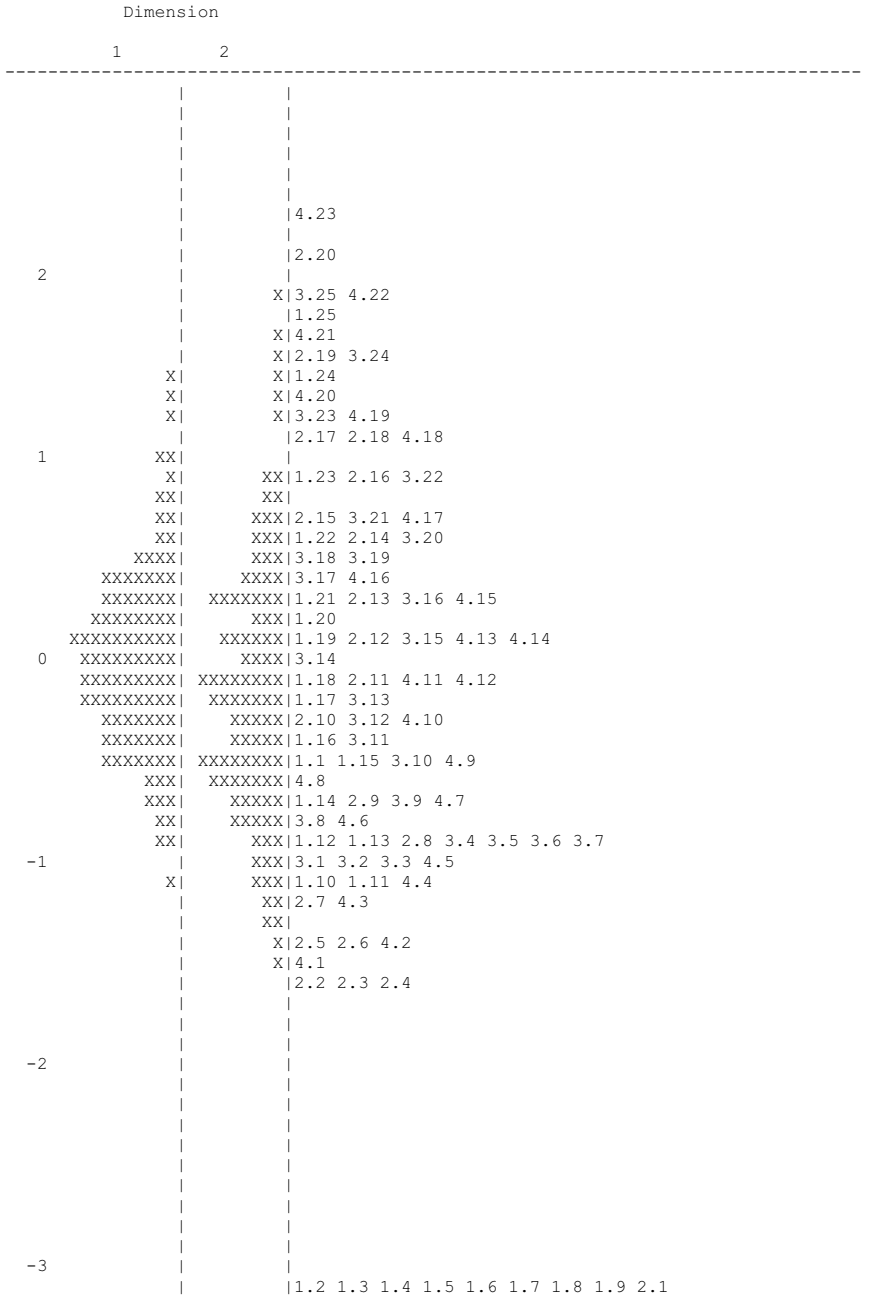

Figure 3:
Map of latent two-dimensional distribution of item thresholds and person abilities

Cross-check of multidimensionality

In order to rule out the possibility that the change in the deviance and the improvement in fit are due to the specific modeling approach adopted rather than being the effect of differences in text types, two other two-dimensional analyses were run.

In the first analysis conversational passage 1 and written passage 1 were modeled to load on the first dimension and conversational passage 2 and written passage 2 to load on the second dimension (Model 3; Figure 4). In the second analysis conversational passage 1 and written passage 2 were modeled to load on the first dimension and conversational passage 2 and written passage 1 to load on the second dimension (Model 4; Figure 5).

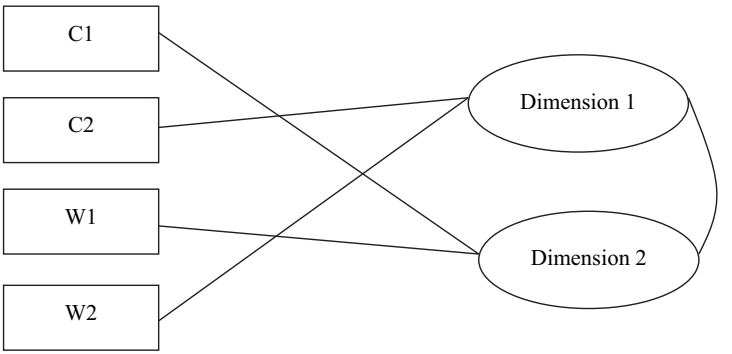

Figure 4:
Graphical representation of Model 3

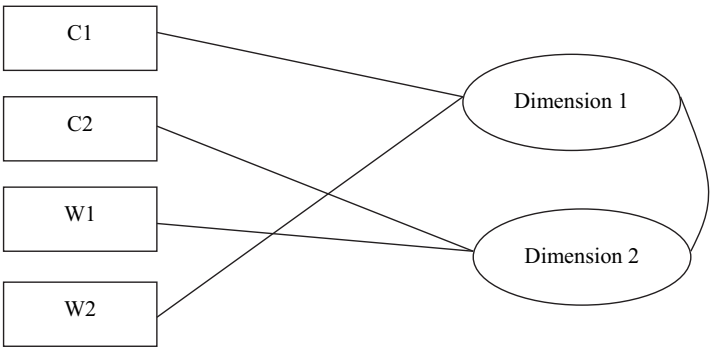

Figure 5:
Graphical representation of Model 4

The statistical results for Models 3 and 4 are shown in Table 4. For Model 3 the deviance is almost equal to the deviance of the unidimensional analysis (cf. Table 3). For Model 4 the deviance is 2014.34, which is much larger than the deviance of 2007.20 for Model 2 where conversational texts and written discourse texts are modeled to load on separate dimensions. Furthermore for Models 3 and 4, the correlation coefficients between the two dimensions are 0.996 and 0.999 and thus considerably higher than for Model 2. All this disconfirms the idea that in Models 3 and 4 the dimensions are distinct. Moreover, when different text types were combined to form one dimension, the model did not converge and iterations stopped because the default maximum number of iterations in ConQuest was reached, while when similar text types were combined the model converged. This is another indication of the validity of the two-dimensional model depicted in Figure 2 above.

Table 5 shows some descriptive statistics for the different analyses which were run in this study. Only for the 'same texts' two-dimensional analysis can we see some noticeable differences in the statistics for the dimensions (Model 2). For the second and third two-dimensional analyses (Models 3 and 4) the statistics of the two dimensions are quite similar and close to the statistics of the unidimensional model.

To sum up, when both written and conversational passages are combined to form one and the same dimension, the two dimensions are in fact identical to what is measured in a unidimensional analysis when all passages are analyzed together. The extremely high correlation between the two dimensions is clear evidence of this fact. However, when similar text types are combined with each other to form two separate dimensions, the fit of the model improves and the correlation between the dimensions drops substantially. This is evidence of the existence of the distinct dimensions that conversational and written discourse C-Test passages define and measure.

Table 4:
Model fit statistics for the multidimensional models

Model	**Deviance**	**Estimated Parameters**	**Correlation Between Dimensions**	**AIC**	**BIC**
Two-Dimensional (Model 3)	2016.81	86	0.996	2188.81	2411.55
Two-Dimensional (Model 4)	2014.34	86	0.999	2186.34	2409.08

Table 5:
Descriptive statistics for four different models

Model	Mean		Variance		Reliability	
	Dimension 1	Dimension 2	Dimension 1	Dimension 2	Dimension 1	Dimension 2
Uni-Dimensional (Model 1)	–0.02	–	0.23	–	0.87	–
Two-Dimensional (Model 2)	0.05	–0.12	0.23	0.58	0.84	0.87
Two-Dimensional (Model 3)	0.00	–0.05	0.23	0.24	0.86	0.87
Two-Dimensional (Model 4)	–0.05	0.01	0.17	0.33	0.87	0.87

Conclusions

The present study showed that MIRT models can be fruitfully applied to test hypothesized structures in language test data. MIRT models can thus be used in the same way as factor-analytic approaches in a confirmatory mode to validate hypothesized test structures. Davison and Skay (1991) even argue in favor of MIRT models as superior alternatives to factor-analytic approaches for confirmatory analyses of test structures. They state that MIRT focuses on variation in task content, while factor analysis focuses on variation in respondents.

A limitation to the study is its small sample size of only 99 subjects. Small sample size results in large sampling error, unstable parameter estimates and model tests with insufficient power. Therefore, the present investigation should be considered to be rather a pilot study, and the conclusions drawn as tentative. Another potential limitation to this study is that the conversational C-Test passages used were drawn from dialogues on ESL websites. One cannot be sure to what extent these dialogues represent authentic spoken discourse as they are manipulated for specific teaching purposes.

Nevertheless, in line with the theoretical arguments and empirical evidence adduced above our analyses suggest that with different types of C-Test texts we can measure different constructs and that the psycholinguistic dimensions hypothesized to be defined by different text types have psychometric grounding too. The better fit of a two-dimensional model compared to a unidimensional model corroborated this hypothesis. The two-dimensional model, when similar text types were combined to load on the same

dimension, fitted significantly better than both a unidimensional model and other two-dimensional models where dissimilar text types were modeled to load on the same dimension. Research with more authentic spoken discourse might make the distinction between two different text types even more pronounced.

Baghaei and Grotjahn (in press) carried out a similar study with a larger sample (n = 200). They investigated whether academic texts and everyday language texts, when converted into C-Tests, can define two separate dimensions of academic and everyday language proficiency. Their findings corroborated that a two-dimensional model of their data, where academic texts and everyday language texts define two separate dimensions, fits significantly better than a unidimensional model where all types of texts are forced to load on a single dimension. As in the present study, the better fit of the two-dimensional model was verified by likelihood ratio test and information criteria. The results of Baghaei and Grotjahn (in press) can be considered as a support for the conclusion arrived in the present investigation regarding the effect of the text features on the construct measured by C-Tests.

We have suggested that our results can also be interpreted that conversational C-Test texts tap into specific aspects of oral/aural ability not accounted for by written discourse C-Test texts. However, more research is required to indicate that this is in fact the case. Future research should include direct tests of listening/speaking and reading/writing skills as criteria to check to what extent these tests correlate with conversational and written-discourse C-Test and whether these tests can be combined with conversational and written discourse C-Tests into common dimensions. All this information would be an important piece of evidence in constructing a comprehensive validity argument for C-Tests (cf. Kane, 2013; Mislevy & Huang, 2007).

Acknowledgment

The Alexander von Humboldt Foundation is gratefully acknowledged for supporting this study by providing a grant for the first author.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Adams, R. J., Wu, M. L., & Carstensen, C. H. (2007). Application of multivariate Rasch models in international large scale educational assessments. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: extensions and applications* (pp. 271-280). New York: Springer.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Alderson, C. J. (2002). Testing proficiency and achievement: principles and practice. In J. A. Coleman, R. Grotjahn & U. Raatz (Eds.), *University language testing and the C-test* (pp. 15-30). Bochum: AKS-Verlag.

Andrich, A. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Arras, U., Eckes, T., & Grotjahn, R. (2002). C-Tests im Rahmen des „Test Deutsch als Fremdsprache“ (TestDaF): Erste Forschungsergebnisse. In R. Grotjahn (Ed.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen* (Vol. 4, pp. 175-209). Bochum: AKS-Verlag.

Baghaei, P. (2008). The effects of the rhetorical organization of texts on the C-Test construct: A Rasch modelling study. *Melbourne Papers in Language Testing, 13*(2), 32-51.

Baghaei, P. (2013). Development and psychometric evaluation of a multidimensional scale of willingness to communicate in a foreign language. *European Journal of Psychology of Education, 28*, 1087-1103.

Baghaei, P., & Grotjahn, R. (in press). The validity of C-Tests as measures of academic and everyday language proficiency: A multidimensional item response modeling study. In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends*. Frankfurt am Main: Lang.

Baghaei, P, Monshi Toussi, M. T., & Boori, A. A. (2009). An investigation into the validity of conversational C-Test as a measure of oral abilities. *Iranian EFL Journal, 4*, 94-109.

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

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

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

Brandt, S. (2012). Robustness of multidimensional analyses against local item dependence. *Psychological Test and Assessment Modeling*, *54*(1), 36-53. Available at: http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2012_20120326/03_Brandt.pdf

Carrell, P. L. (1984). The effects of rhetorical organization on ESL readers. *TESOL Quarterly*, *18*(3), 441-469.

Chapelle, C. A., & Abraham, R. G. (1990). Cloze method: what difference does it make? *Language Testing, 7*(2), 121-146.

Coleman, J. A. (1994). Profiling the advanced language learner: the C-Test in British further and higher education. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 2, pp. 217-237). Bochum: Brockmeyer. Available at: http://www.c-test.de/deutsch/index.php?lang=de&section=originalia.

Daller, H., & Grotjahn R. (1999). The language proficiency of Turkish returnees from Germany: An empirical investigation of academic and everyday language proficiency. *Language, Culture and Curriculum*, *12*(2), 156-172.

Daller, H., & Phelan, D. (2006). The C-test and TOEIC® as measures of students' progress in intensive short courses in EFL. In R. Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, empirical research, applications* (pp. 101-119). Frankfurt am Main: Lang.

Davison, M. L., & Skay, C. L. (1991). Multidimensional scaling and factor models of test and item responses. *Psychological Bulletin, 110*, 551-556.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-Test amongst Hungarian EFL learners. *Language Testing, 9*(2), 187-206.

Eckes, T. (2006). Rasch-Modelle zur C-Test-Skalierung. In R. Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen/The C-Test: Theory, empirical research, applications* (pp. 1-44). Frankfurt am Main: Lang.

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling, 53*(4), 414-439. Available at: http://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2011_20111217/02_eckes.pdf

Eckes, T. (in press). Die onDaF–TestDaF-Vergleichsstudie: Wie gut sagen Ergebnisse im onDaF Erfolg oder Misserfolg beim TestDaF vorher? In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends*. Frankfurt am Main: Lang.

Eckes, T., & Baghaei, P. (in press). Using testlet response theory to examine local dependency in C-Tests. Applied Measurement in Education.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*(3), 290-325.

Embretson (Whitely), S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

Embretson (Whitely), S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson (Whitely), S. E. (1985). Introduction to the problem of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 3-17). Orlando, Fla.: Academic Press.

Embretson, S. E. (2007). Mixed Rasch models for measurement in cognitive psychology. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: extensions and applications* (pp. 235-253). New York: Springer.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77-151). Cambridge: Cambridge University Press.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Grotjahn, R. (1992). Der C-Test im Französischen. Quantitative Analysen. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Vol. 1, pp. 205–255). Bochum: Brockmeyer. Available at: http://www.c-test.de/deutsch/index.php?lang=de&section=originalia

Grotjahn, R. (2013). Cloze test. In M. Byram & A. Hu (Eds.), *Routledge encyclopedia of language teaching and learning* (2nd ed., pp. 121-122). London: Routledge.

Grotjahn, R. (in press). The C-Test bibliography: Version January 2014. In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends*. Frankfurt am Main: Lang.

Harsch, C., & Hartig, J. (2010). Empirische und inhaltliche Analyse lokaler Abhängigkeiten im C-Test. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 193-204). Frankfurt am Main: Lang.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology, 216*(2), 89-101.

Hartig, J., & Höhler, J. (2010). Modellierung von Kompetenzen mit mehrdimensionalen IRT-Modellen: *Projekt MIRT*. In E. Klieme, D. Leutner & M. Kenk (Eds.), *Kompetenzmodellierung: Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (pp. 189-198). Weinheim: Beltz. Available at: http://www.pedocs.de/frontdoor.php?source_opus=3324&la=de.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Höhler, J., Hartig, J., & Goldhammer, F. (2010). Modeling the multidimensional structure of students' foreign language competence within and between classrooms. *Psychological Test and Assessment Modeling, 52*(3), 323-340. Available at: http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2010_20100928/07_Hoehler.pdf.

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research, 34*, 245-268.

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

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(4), 331-358.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie/Journal of Psychology, 216*(2), 61-73.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model – some critical suggestions on traditional approaches. *International Journal of Testing, 5*(4), 377-394.

Lane, S., & Stone, C. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-431). Westport, CT: American Council on Education and Praeger Publishers.

Lin, T. H., & Dayton, M. C. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*(3), 249-264.

Linacre, J. M. (1989). Many-facet Rasch measurement. Chicago: MESA Press.

Masters, J. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

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

Mislevy, R. J., & Huang, Chun-Wei. (2007). Measurement models as narrative structures. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 16-35). Berlin: Springer.

Moosbrugger, H., & Müller, H. (1982). A classical latent additive test model (CLA model). *The German Journal of Psychology, 6*(2), 145-149.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research 1960. (Expanded edition, Chicago: The University of Chicago Press, 1980).

Reckase, M. D. (1997). The past and the future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25-36.

Ridden, G. M. (1982). York notes on Shakespeare's sonnets. Harlow, England: York Press/Longman.

Santelices, M. V., & Caspary, K. (2009). Development of a multidimensional measure of academic engagement. *Journal of Applied Measurement, 10*(4), 371-393.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Shohamy, E. (1982). Predicting speaking proficiency from Cloze tests: Theoretical and practical considerations for tests substitution. *Applied Linguistics, 3*(2), 161-171.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: the effect of text and question type. *Language Testing*, *8*(1), 23-40.

Sigott, G. (2004). *Towards identifying the C-Test construct*. Frankfurt am Main: Lang.

Walter, O. (2005). *Kompetenzmessung in den PISA-Studien. Simulationen zur Schätzung von Verteilungsparametern und Reliabilitäten.* Lengerich: Pabst Science Publishers.

Wang, W.-C., Cheng, Y.-Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement, 65*(1), 5-27.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126-149.

Wilson, M., & Moore, S. (2011). Building out a measurement model to incorporate complexities of testing in the language domain. *Language Testing, 28*(4), 441-462.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2: Generalised item response modelling software*. Camberwell: Australian Council for Educational Research.

Appendix

CONVERSATIONAL C-TEST PASSAGE 1 (C1)

Alan: Betty, is it possible to borrow your notes? I'll return them tomorrow.

Betty: Sorry, b_____ I usu_____ go t_____ the café_____ and rev_____ them. S_____, how ab_____ copying th_____ over i_____the lib_____?

Alan: Okay, I th_____ I've g_____ enough co_____ for t_____ machines. Yo_____a life_____, Betty!

Betty: N_____ problem. B_____ I do_____ understand w_____ you ne_____ my no_____, Alan; y_____ haven't mis_____ any cla_____.

Alan: Weekday mornings, I'm a cashier at a coffee shop downtown. After work, I come directly to school, and, boy, am I beat!

Betty: Wow, you're probably exhausted!

WRITTEN C-TEST PASSAGE 1 (W1)

There can be no doubt that Shakespeare is generally regarded as the greatest playwright who ever lived. Throughout t_____ world h_____ plays cont_____ to b_____ performed, a_____ memorable li_____ from th_____ have sli_____ almost unno_____ into ever_____ use. Shakespeare's dram_____ genius, alth_____ not wid_____ recognized b_____ his contemp_____, was, how_____, acclaimed we_____ before h_____ was recog_____ as a gr_____ poet. Alth_____ the dat_____ of mu_____ of Shakespeare's wo_____ is unce_____ and it may not be the case that he turned from poems to plays quite in the manner suggested by Halliday, it is true that his first published work was Venus and Adonis, published in 1593 and reprinted fifteen times before 1640.

CONVERSATIONAL C-TEST PASSAGE 2 (C2)

Salesman: Hi, are you being helped?

Karen: No, I_____ not. I_____ interested i_____ some sca_____.

Salesman: All o_____ scarves a_____ in th_____ section. Wh_____ do y_____ think o_____ this o_____ here? I_____ made o_____ silk.

Karen: Hm, i_____ looks ni_____, but I_____ like t_____ have some_____ warm f_____ the win_____.

Salesman: Maybe y_____ would li_____ a he_____ wool sc_____. How ab_____ this one?

Karen: I think that's what I want. How much is it?

Salesman: It's...seventy-five dollars plus tax.

WRITTEN C-TEST PASSAGE 2 (W2)

In January there came bitterly hard weather. The ea_____ was li_____ iron, a_____ nothing co_____ be do_____ in t_____ fields. Ma_____ meetings we_____ held i_____ the b_____ barn, a_____ the pi_____ occupied thems_____ with plan_____ out t_____ work o_____ the com_____ season. I_____ had co_____ to b_____ accepted th_____ the pi_____, who we_____ manifestly clev_____ than t_____ other animals, should decide all questions of farm policy, though their decisions had to be ratified by a majority vote. This arrangement would have worked well enough if it had not been for the disputes between Snowball and Napoleon.