Harsch, Claudia; Hartig, Johannes

Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills

Language testing (2015), S. 1-21

Article

LANGUAGE TESTING

Language Testing
1–21

DOI: 10.1177/0265532215594642
ltj.sagepub.com
SAGE

Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills

Claudia Harsch
University of Warwick, UK

Johannes Hartig
German Institute for International Educational Research (DIPF), Germany

Abstract

Placement and screening tests serve important functions, not only with regard to placing learners at appropriate levels of language courses but also with a view to maximizing the effectiveness of administering test batteries. We examined two widely reported formats suitable for these purposes, the discrete decontextualized Yes/No vocabulary test and the embedded contextualized C-test format, in order to determine which format can explain more variance in measures of listening and reading comprehension. Our data stem from a large-scale assessment with over 3000 students in the German secondary educational context; the four measures relevant to our study were administered to a subsample of 559 students. Using regression analysis on observed scores and SEM on a latent level, we found that the C-test outperforms the Yes/No format in both methodological approaches. The contextualized nature of the C-test seems to be able to explain large amounts of variance in measures of receptive language skills. The C-test, being a reliable, economical and robust measure, appears to be an ideal candidate for placement and screening purposes. In a side-line of our study, we also explored different scoring approaches for the Yes–No format. We found that using the hit rate and the false-alarm rate as two separate indicators yielded the most reliable results. These indicators can be interpreted as measures for vocabulary breadth and as guessing factors respectively, and they allow controlling for guessing.

Keywords

C-test, placement test, predictive power, receptive language skills, structural equation modelling, Yes/No vocabulary test

Corresponding author:
Claudia Harsch, The Centre for Applied Linguistics, The University of Warwick, Coventry, CV4 7AL, UK.
Email: C.Harsch@warwick.ac.uk

Placement and screening tests serve important functions, not only with regard to placing learners at appropriate levels of language courses but also with a view to maximizing the effectiveness of administering test batteries. With the advent of computer-administered tests (CAT), adaptive testing (e.g., Frey & Seitz, 2009) and multi-stage testing (e.g., Yan, von Davier, & Lewis, 2014) are gaining importance. In the emerging field of CAT, placement tests can increase the efficiency of CAT procedures (e.g., Frey & Seitz, 2009). When administered before the actual CAT battery, they can help determine the starting level at which the adaptive test begins. As stated by, for example, Bachman & Palmer (1996) or Read (2000), the test purpose determines all future decisions. Placement or screening purposes require a format that has been reported as being a reliable predictor for the skills targeted in the main test, is simple and quick in administration and scoring, and makes few demands on the test takers while covering as many items as possible.

The literature reports two kinds of tests in particular as being feasible for placement and screening purposes, that is, vocabulary tests and C-tests. Different kinds of vocabulary tests have been reported as being valid predictors specifically for reading skills (e.g., Alderson, 2005; Alderson & Huhta, 2005; Elder & von Randow, 2008; Harrington & Carey, 2009; Read, 2000) and as predictors of general language proficiency, for example as controls in experimental studies (e.g., P. M. Meara, pers. com., December 2014; or Milton, 2009). Vocabulary tests have also been used in CAT procedures (Laufer & Goldstein, 2004). With regard to C-tests, most studies report them as being predictors of general language proficiency (e.g., Eckes & Grotjahn, 2006; Grotjahn, 2002; Klein-Braley, 1985; Klein-Braley & Raatz, 1984; Harsch & Schröder et al., 2007), but some researchers discuss them as measures for reading and/or vocabulary (e.g., Chapelle, 1994; Eckes & Grotjahn, 2006; Read, 2000). The relationship between different vocabulary measures and C-tests has been investigated in the context of validating vocabulary measures (e.g., Chapelle, 1994), and in the context of examining the applicability of C-tests for vocabulary assessment (Karimi, 2011). However, vocabulary measures and C-tests have not yet been directly compared in their capacity to predict reading skills. Much less is known about the predictive power of vocabulary measures and C-tests for listening skills, which is why our study focuses on examining the predictive power of vocabulary test and C-tests for the receptive skills of both listening and reading.

Using Read's (2000) and Read and Chapelle's (2001) framework for assessing vocabulary as backdrop, specifically the dimensions 'discrete vs. embedded' and 'decontextualized vs. contextualized', our study focuses on two formats operationalizing the opposing ends of these two dimensions. We investigate whether a discrete, decontextualized approach in the format of a Yes/No vocabulary test with pseudowords to control for guessing (X-Lex test by Meara, 2005a, b) or the embedded, contextualized C-test format (items from the DESI study, see below, Harsch & Schröder et al., 2007, 2008) can explain more variance in test scores for reading and listening comprehension.

We first discuss the literature on the constructs of the two formats and on their validity for placement purposes, before we contrast the formats' reported explanatory power for the receptive skills.

C-test: Construct and use as placement test

The C-test format plays a prominent role in language testing (Eckes & Grotjahn, 2006). A C-test usually consists of four or five short texts that are coherent in themselves. Within each of these texts, between 20 and 25 words are chosen either randomly (e.g., by choosing every second word) or according to certain construct-related principles. The second half of these words is deleted, and the test takers are to complement the missing halves (see, e.g., Grotjahn, 2002 for construction principles). The C-test is reported as being an economical, objective, easy to administer and to score, reliable and valid format for placement and selection purposes (e.g., Eckes & Grotjahn, 2006; Grotjahn, 2002), which makes it a suitable format for our study.

Having placed the C-test at one end of a possible spectrum of different means to test vocabulary, this is not to say that we would regard the C-test as a measure of vocabulary, as for example discussed by Chapelle (1994), Eckes & Grotjahn (2006) or Read (2000, pp. 101–116). Rather, we agree with the vast body of research reporting a complex construct for the C-test. A C-test is based on the principle of reduced redundancies, the assumption being that the more proficient a learner is, the more gaps they can solve by drawing on their automated language skills (e.g., Coleman, 1996; Klein-Braley, 1997). Solving C-tests elicits a complex array of language skills and linguistic knowledge, as well as (meta-)cognitive strategies in an integrative way (e.g., Grotjahn, 2002; Hastings, 2002; Klein-Braley, 1985; Sigott, 2004; Stemmer, 1991). The C-test construct is partly influenced by test-taker characteristics such as L1 background, closeness of L1 and L2, and language proficiency (see, e.g., the overview in Eckes & Grotjahn, 2006, pp. 293–294; Reichert, Brunner, & Martin, 2014). Hence, a C-test may be measuring different skills for different learners, a phenomenon which Sigott (2004, 2006) calls a 'fluid' construct.

In sum, the C-test is widely reported as being a reliable and valid measure of general language proficiency, fit for the purposes of placement or screening, since it allows for the sampling of a large number of items (i.e., the deleted word halves) in a relatively short time (e.g., Eckes, 2014; Eckes & Grotjahn, 2002, 2006; Klein-Braley, 1985; Klein-Braley & Raatz, 1984; Harsch & Schröder et al., 2007).

Yes/No test: Construct and use as placement test

The literature distinguishes between approaches to measure vocabulary breadth and those targeting vocabulary depth (for an overview of different measures of vocabulary, see, e.g., Read, 2000, or Milton, 2009; for an overview of research into vocabulary measures, see Read, 2013). Measures of vocabulary breadth are reported as being valid, reliable and suitable for placement and screening purposes (e.g., Elder & von Randow, 2008; Milton, 2009, p. 171; Read, 2000, p. 162). One prominent and well-researched format to measure vocabulary breadth is the so-called Yes/No format, where isolated words are presented, their selection based on frequency lists (e.g., Meara, 2005b; Mochida & Harrington, 2006; Pellicer-Sánchez & Schmitt, 2012). Test takers are asked to indicate if they know the meaning of the words. The assumption behind this format is 'that the more frequent a word is in a language … the more easily, and the earlier, it is likely to be learned' (Milton,

2007, p. 48). In order to control and correct for guessing, pseudowords are used (i.e., words that resemble the morphological rules of real words but do not exist). Examples for this format are the Eurocentres Vocabulary Size test used for placement purposes (Meara & Jones, 1990), the X-Lex (Meara, 2005a) used as indicator for learners' 'overall proficiency levels' (Meara, 2005b, p. 21) or the vocabulary placement test used in the online diagnostic test DIALANG (Alderson, 2005; Alderson & Huhta, 2005).

Similarly to the above-mentioned fluidity of the C-test construct, research on the Yes/No format also suggests that test-takers' characteristics, such as general proficiency level or closeness of L1 and L2, have an influence on the scores for items on different frequency bands and on the answer behaviour for pseudowords (e.g., Huibregtse, Admiraal, & Meara, 2002; Milton, 2007; Read, 2000). When it comes to scoring the Yes/No format, there are different ways to take into account test-takers' responses to the pseudowords (e.g., Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001; Huibregtse et al., 2002; Mochida & Harrington, 2006; Pellicer-Sánchez & Schmitt, 2012). So far, no clear advantage of any one method has been reported. Thus, in a side-line of our study, we will also explore a suitable scoring method for our sample and purposes.

In brief, the Yes/No format, in a striking analogy to the C-test format discussed above, has also been reported as being suitable, reliable and valid for placement and screening purposes. It allows for the sampling of a high number of items in a short time, is easy and economical to administer and score, and does not place undue strain on test takers (e.g., Beeckmans et al., 2001; Harrington & Carey, 2009; Milton, 2009; Nation, 1990; Pellicer-Sanchez & Schmitt, 2012; Read, 2000). Since no study has yet directly compared these two formats in their potential as placement or screening tests, our study is addressing this gap in the literature.

Predictive power

When using tests for placement or screening purposes, it is important to examine the tests' predictive power for the domain in which score interpretations are to be used. A test being used as an indicator of general proficiency should yield close relationships with other tests that measure all four skills. A test being employed as indicator for reading is expected to share a large amount of variance with other measures of reading comprehension. For the purpose of the research we report here, we now look at findings from studies which investigated the predictive power of C-tests and Yes/No vocabulary tests.

Yes/No vocabulary tests are generally regarded as indicators of 'receptive' vocabulary knowledge (e.g., Mochida & Harrington, 2006). As is the case for all tests of vocabulary breadth (e.g., Milton, 2009; Read, 2000, p. 162), the underlying assumption is that if a learner cannot recognize an item as a word in a specific language, it is unlikely that the learner can do anything else with the word (P. M. Meara, pers. com., December 2014). This fundamental assumption is one reason why we selected a vocabulary measure focusing on 'receptive' vocabulary knowledge for the study reported here. With regard to listening and reading comprehension, Nation (2006) described the necessary size of vocabulary to understand certain texts and inputs. For text comprehension to take place, he states that a reader's vocabulary knowledge needs to cover 98% of a text's vocabulary. This emphasizes the important role vocabulary plays in comprehension processes. This is not to say that we

ignore the intricate relationship between vocabulary and grammar when processing texts, or regard vocabulary more important than grammar, an area where research has also not yet yielded clear results (see, e.g., Alderson & Kremmel, 2013 or Shiotus & Weir, 2007). However, introducing yet another variable, that is, grammar, into our study would result in too complex a design. Hence, we focus only on vocabulary in this study.

Different vocabulary measures were examined by Qian (2008) for their predictive power on reading test scores. His findings showed little difference between isolated word list formats and contextualized vocabulary tests in their predictive power for reading comprehension. Notwithstanding the close relationship between the Yes/No format (and other measures of vocabulary breadth) and reading/receptive skills, many researchers recommend and use the Yes/No test as a general placement test (e.g., Alderson, 2005; Alderson & Huhta, 2005; Meara, 2005b, p. 21; Meara & Jones, 1990; Read, 2000, p. 162), as it also appears to be a reliable indicator for general language proficiency (e.g., Milton 2009, p. 171). While we aim at investigating the suitability of this format as a placement or screening test, we will restrict our interpretations to the receptive skills of listening and reading for the study reported here. Thus, we aim at extending the well-researched area of reading to the realm of listening comprehension since not many studies have been reported in this area.

C-tests on the other hand have been reported to be valid predictors of general language proficiency since they tend to show high correlations with other tests measuring the four skills and vocabulary knowledge (see, e.g., Eckes & Grotjahn, 2006 for an overview of correlational studies). In a recent study, Eckes (2014) investigated the predictive power of a C-test used in an online placement test (onDaF) for the TestDaF, a test of general language proficiency used as university language entrance test. Eckes found 'consistently high … rates of agreement between examinee level assignments … provided by onDaF and TestDaF' (2014, p. 137), so that the former could serve as a quick screening test to check whether test takers are 'ready' to take the TestDaF.

When comparing C-tests and vocabulary measures, most studies found a close relationship, which is why some researchers discussed C-tests as being contextualized measures of vocabulary (see above). When Chapelle (1994) examined the usefulness of C-tests for vocabulary research, she concluded that more clarity was required regarding the C-test construct and the nature of the C-test items in order for it to be a useful tool for research. More than a decade later, drawing on research by, for example, Sigott (2004, 2006) and many others, Eckes and Grotjahn (2006) explained high correlations reported in the literature between C-tests and vocabulary measures through the assumption that the C-test is tapping into the test-takers' mental lexicon, amongst other skills (pp. 298–299). Thus, although the relationship between C-tests and vocabulary measures has been examined to a certain extent, no attempt is reported in the literature to compare directly C-tests and Yes/No vocabulary tests in their power to predict performance on listening and reading comprehension tests.

Aims and research questions

Based on the gaps identified in the literature above, we aim to compare the predictive power of a C-test and a Yes/No vocabulary test on measures of listening and reading

comprehension. First, we examine the relationship between all four measures. We then analyse what amount of variance in measures of reading and listening comprehension can be explained by the test scores achieved on the C-test and the Yes/No vocabulary test, in order to decide which of the two is the better predictor for our context and sample. While we acknowledge the influence of the test-takers' characteristics on the performance on the predictor formats as described above, our sample is a rather homogeneous group of test takers (see below). Hence we did not differentiate the test-takers' characteristics in our analyses. We aim to answer the following research questions:

1. What is the relationship between the different measures gained from the tests for listening and reading comprehension, C-test and the Yes/No vocabulary test?
2. Which of the two measures, C-test or Yes/No vocabulary test, is the more effective predictor for listening and reading comprehension?

Both the relationships between the different measures, as well as the predictive power of the C-test and the vocabulary test will be examined on the level of observed test scores and on a latent level, using structural equation modelling (SEM; e.g., Kunnan, 1998; Purpura, 1998). The observed scores are relevant for contexts in which a decision is based on individual test-takers' scores, whereas SEM allows a better insight into the relationships of the underlying latent test constructs, since the relations between the latent variables are corrected for measurement error.

Methods

Context

The study we report here is situated within the context of a large-scale assessment aimed at evaluating the German Educational Standards that were introduced in 2003 (KMK, 2003). The test instruments were developed and administered by the Institute for Educational Quality Improvement, Humboldt-University Berlin. For a detailed account of the test development, see Rupp, Vock, Harsch, and Köller (2008), and for an account of calibrating the item pool and aligning it to the CEFR (Council of Europe, 2001), see Harsch, Pant & Köller (2010). In 2008, a large pool of piloted items was administered to a nationally representative sample of ninth and tenth graders (14 to 16 years old) attending the middle and higher tracks of secondary schools in Germany (Köller, Knigge, & Tesch, 2010). Overall, 3,404 students were tested in a multi-matrix design (i.e., not every test taker worked on all test items), but the design ensured that all items were linked. This linkage allowed for the estimation of test scores for all students on a common scale through scaling the data using item response theory (IRT) modelling. Specifically, the data in the main study were analysed using the Rasch model (Köller et al., 2010). The items were organized in blocks, and the blocks were distributed according to the above-mentioned multi-matrix design across a total of 24 test booklets. The four instruments we employed in our study were administered within this large-scale assessment, along with a range of other instruments. In what follows, we limit our description to the subsample and the four instruments relevant to our study.

Sample

The student sample consisted of 14–16-year-old students attending the ninth and tenth grades of the middle and upper tracks of the German three-tier secondary school system. The students' English proficiency was at an intermediate level, and the majority of the students had German as their first language, German also being the language of schooling (Köller et al., 2010). As mentioned above, a total of 3404 students were sampled from 148 intact classes for the large-scale assessment. The subsample reported in this paper encompassed 559 students from 24 intact classes, working on the test booklets containing the four instruments examined in this study.

Instruments

For the study reported here, we used four test instruments targeting reading comprehension, listening comprehension, a vocabulary test using items from the X-Lex test and a C-test. The test instruments are described in more detail below. All four instruments were administered in the aforementioned large-scale assessment. The listening test was administered in 22 booklets (n = 3,102) and the reading test in 16 booklets (n = 2,279) in the above-mentioned multi-matrix design. The number of responses per item ranged from 244 to 886 for listening and from 330 to 727 for reading. Both the C-test and the X-Lex items were administered in four of the 24 booklets to the above mentioned subsample of n = 559 students. All four booklets containing C-test and X-Lex items also contained a subset of listening and reading items. This booklet design will be described in more detail below.

C-tests. The C-tests in this study were selected from a larger pool of 12 texts developed in the context of an earlier large-scale study assessing ninth graders' proficiency in English as foreign language and German as first language (DESI study; see Harsch & Schröder, 2007, 2008 for instruments and instructions). The texts were validated for the same learner population from which the sample of our study is drawn. For the study reported here, four texts with 25 items each were selected from the DESI pool, covering a range of difficulties (from beginner to upper levels), with a slight focus on the intermediate to upper levels to match the proficiency level of our subsample. The four texts in the study reported here were administered in two different versions, which contained the texts in different orders to control for positioning effects and to prevent students from copying the answers from their neighbours. The C-test was employed in four test booklets. The students had 20 minutes for this part of the test.

Vocabulary measure. We used 120 vocabulary items from Meara's X-Lex test (2005a): 20 items randomly selected per frequency band for the first five bands containing 1000 words each, resulting in 100 vocabulary items; and 20 pseudowords used in X-Lex, also randomly selected. Since our test-taker sample was generally of intermediate proficiency (Köller et al., 2010), the X-Lex should be suitable for this learner group (Meara, 2005b). In the study reported here, all items were randomly ordered and administered in two different versions, which contained the items in reverse order to control for positioning

Table 1. Distribution of listening blocks across booklets.

Booklet	Listening blocks	
A	L1	L2
B	L1	L2
C	L3	L4
D	L3	L4

effects and to prevent students from copying. The instruction we used in this administration is presented in Appendix A. The X-Lex items were employed in the same four booklets as the C-test. Again, the students had 20 minutes for the X-Lex items.

Listening test. The listening comprehension test administered in the above-mentioned large-scale assessment operationalized the German educational standards (for details on the constructs, instruments and instructions, see Rupp et al., 2008). The test items spanned CEFR levels A2 to B2 and covered a range of formats (e.g., multiple choice, table filling, short-answer questions, matching). Overall, a total of 41 tasks with 194 items was employed in 22 booklets. The booklet design for the total sample (n = 3404) ensured that all items could be linked to allow for the construction of test scores on a common scale by means of IRT scaling, as outlined above.

The subsample used for the present study, $n = 559$ students, worked on those listening items contained in the above-mentioned four test booklets which also held the C-test and the X-Lex items. Within these four booklets, four blocks of listening tests with a total of 132 items in 20 tasks with four or more items, and three single-item tasks were employed. The students had 20 minutes to work on each block. The four blocks of items (L1 to L4, five tasks in each block) were distributed across the four booklets (A to D) as shown in Table 1.

Hence, the two booklets A/B shared two blocks of items (L1/L2), while booklets C/D shared a set of two different blocks (L3/L4). However, there is no overlap of items between booklets A/B on the one hand, and booklets C/D on the other hand. This occurrence of non-overlapping blocks of listening items has consequences for our analysis, which will be discussed below in the data analysis section.

Reading test. The complete reading comprehension test used for the large-scale assessment outlined above consisted of 43 tasks with 214 items in 16 booklets. The items covered the same range of CEFR levels and formats as the listening test (e.g., multiple choice, table filling, short-answer questions, matching; for more details on the instruments and instructions, see Rupp et al., 2008). Similar to the listening test, linkage of all items was ensured for the total sample, so that test scores could be constructed on a common scale by means of IRT scaling.

The subsample of the present study responded to three blocks of reading items, which were non-overlapping within the above-mentioned four test booklets (i.e., not all three reading blocks were contained in each of the four booklets). Again, the consequences for our analyses will be discussed below. The three relevant reading blocks contained 80

items in 13 tasks with four or more items, and one single-item task. As for the listening tests, the students had 20 minutes to work on each block.

Data preparation

For the regression analyses, we needed to calculate overall observed scores for each of the four test instruments. Additionally, to conduct latent regression, multiple sub-scores (i.e., separate scores calculated from different parts of each of the four tests) were calculated for each instrument to be used as indicators in structural equation modelling (SEM). In what follows, we outline the calculation of the overall scores and sub-scores for each of the instruments in turn.

C-test. For the C-test, the overall scores were constructed as the percentage of correct responses for all 100 items (four texts with 25 gaps each). The overall scores had a reliability of $\alpha = .95$. With regard to the multiple sub-scores, we calculated the percentage of correct responses separately for each of the four texts, resulting in four sub-scores for the C-test to be used in SEM.

Vocabulary measure. For the X-Lex, two overall scores were constructed. First, we calculated the hit rate (HR) across the 100 vocabulary items, that is, the percentage of correct responses for the real words. Second, we calculated the false-alarm rate (FAR) across the 20 pseudowords, that is, the percentage of students indicating that they knew this pseudoword. The HR had a reliability of $\alpha = .94$, the FAR of $\alpha = .76$. We were using HR and FAR as separate overall observed scores, as they capture the essential diagnostic information provided by X-Lex. A number of suggestions have been made to integrate both into one common score, for example by calculating the difference between HR and FAR (Meara, 2005b) or an index based on signal detection theory, as suggested by Beeckmans et al. (2001) and Huibregtse et al. (2002). We tested these alternative scores as predictors in additional analyses, but none of them had a better predictive power than HR and FAR as two separate variables. To obtain multiple sub-scores as indicators for SEM, the HR was calculated separately for each of the five frequency bands. In addition to these five sub-scores, we also employed the FAR, resulting in six sub-scores for X-Lex to be used in SEM.

Listening test. For the listening test, the overall observed scores were constructed by means of IRT scaling of the data from the total sample. This allowed the construction of scores on a common scale despite the above-mentioned non-overlapping sets of items answered by our subsample. The Rasch model was applied as measurement model, and weighted likelihood estimators (WLEs; Warm, 1989) were used as individual ability scores (i.e., our overall observed scores). WLE reliability for the listening scores was .88. IRT scaling and score estimation were conducted using the R-package TAM (Kiefer, Robitzsch, & Wu, 2014). This procedure is consistent with the routines used to analyse the listening test in the main study to evaluate the German Educational Standards (Köller et al., 2010). To obtain multiple indicators for SEM, we used the items which were contained in the above-mentioned four test booklets (A to D).

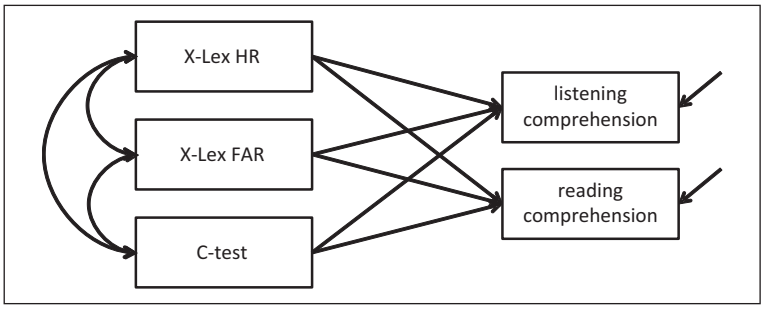

Figure 1. Complete regression model (model 3) used for the prediction of observed scores in listening and reading. In model 1, only the X-Lex scores are used as predictors, and in model 2 only the C-Test score is used.

These items were actually answered by our subsample. Percent-correct scores were calculated for the 20 tasks containing four or more items, resulting in 20 sub-scores to be used as indicator variables for SEM. The three single-item tasks were not included in these indicators.

Reading test. For the reading test, the same procedure as for the listening test was applied. As overall observed scores, WLEs from Rasch scaling of the total sample were used. WLE reliability for the reading score was .86. As sub-scores to be used as indicators for SEM, percent-correct scores were calculated for the 13 tasks contained in the above-mentioned four booklets administered to our subsample. These tasks contained more than four items each; one single-item task was excluded.

Results

Data analysis

To answer research question 1 (i.e., to inspect relationships between tests on a descriptive basis), correlations both on the observed and the latent levels were conducted: we conducted correlational analyses between the observed overall scores and between the latent variables. To examine research question 2 (i.e., the predictive power of the X-Lex and the C-test on measures of reading and listening comprehension), we again looked at the observed and the latent levels.

First, on the level of observed test scores, linear regression analyses with the listening and reading scores as dependent variables were conducted. For both dependent variables, two models using only the X-Lex scores (regression model 1) and only the C-test score (model 2) as predictors were calculated; a third model used the X-Lex and C-test scores simultaneously (model 3). Figure 1 depicts the third regression model containing all predictors.

The regression models examine the effectiveness of the observed Yes/No vocabulary and C-test scores as predictors for listening and reading comprehension. By comparing the three models, we can find out which of the two tests (X-Lex and C-Test) is the stronger predictor for listening and reading scores, and to what extent each of the two

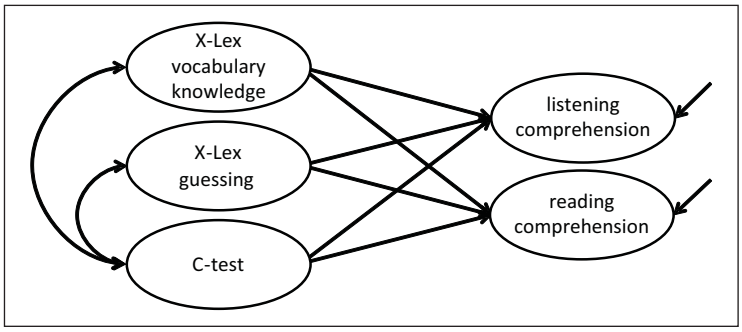

Figure 2. SEM predictive model containing all three latent predictor variables. Indicator variables and measurement models are not depicted for the sake of comprehensibility.

tests can account for variation in listening and reading scores that cannot be explained by the other test. The regression models using the observed scores show the strength of relations and the amount of variance that would be of practical interest in an applied setting (e.g., when students are screened using one test and inferences are made about the performance students are likely to show in another test).

Second, to examine research question 2 for the latent level, we conducted SEM analyses. Here, the predictive power of the X-Lex and the C-test is examined at the level of the underlying constructs. The constructs are represented by so called latent variables. To calculate these latent variables, multiple indicators are required for each construct. Technically, the latent variables within the structural equation models (SEM) can be regarded as dimensions representing the common variation of all indicators for one construct. We gained these indicators by means of the above-mentioned sub-scores, which were based on the test scores of our four measures. In SEM, the relations between the different constructs measured by the different tests are corrected for measurement error. This allows inferences about how much variance the different tests share on the level of the underlying constructs (i.e., to what extent the same abilities are required for a successful performance).

We analysed latent regression models with a structure parallel to the regression analyses described above for the observed scores. However, we now used latent variables representing the constructs assessed by X-Lex and C-Test as predictors for the latent variables representing listening and reading comprehension. As for the observed scores, three different latent regression models were analysed. In model 1, only the X-Lex dimensions were used as predictors, in model 2 only the C-Test dimension was used. Model 3 included all latent predictor variables, as is shown in Figure 2.

In preparation of the SEM analyses, confirmatory factor analyses (CFA) were conducted separately for each construct. This is necessary to see whether the indicators used for each construct actually assess the assumed common dimension(s). The CFA are discussed in the next section.

All analyses were conducted using Mplus 7.11 (Muthén & Muthén, 2012). Since the students were sampled from intact classrooms, standard errors and model fit indices needed to be corrected for the effects of the sample structure. This was done using the pseudo maximum likelihood estimation method implemented in Mplus (Asparouhov & Muthén, 2005).

Confirmatory factor analyses

As indicated above, CFAs were conducted separately for each of the four tests, in preparation for the SEM analyses. For the C-test and the listening and reading items, unidimensional models were tested. The multiple indicators described above for each test were used as indicators for one underlying latent dimension. For listening, unidimensionality was also tested separately for the above-mentioned two sets of non-overlapping items. The analyses for listening and reading were based on those items included in the four booklets which also contained X-Lex and C-test. Since these listening and reading items were also included in other test booklets, we included all students working on these items. (This explains the varying numbers of participants in Table 2.)

The CFA for the X-Lex tested a different and somewhat more complex model, since correct responses to the vocabulary items (hits) and positive responses to the pseudowords (false alarms) were not assumed to constitute one homogeneous dimension. The model contained one common factor for the vocabulary items, using the scores for the five bands as indicators. Additionally, a factor for guessing behaviour was defined, which used the pseudowords score (i.e., the false-alarm rate) as indicator and additionally had loadings on all five vocabulary item indicators; this factor served to control for guessing. The vocabulary factor and the pseudoword factor were restricted to be uncorrelated for model identification purposes.

Subsequent to the CFAs for the separate tests, a CFA for the complete data set was conducted. The CFA models for the separate tests were used as measurement models, and the latent correlations between the constructs were freely estimated. For the evaluation of model fit, we used cut-off criteria recommended by Hu and Bentler (1999). The comparative fit index (CFI) and Tucker-Lewis index (TLI) should be at least .95, the standardized root mean square residual (SRMR) should not exceed .09. Additionally, the root mean square error of approximation (RMSEA) should not exceed .05 (MacCallum, Browne, & Sugawara, 1996).

Table 2 provides an overview of the resulting fit statistics. Additional to the fit statistics used to determine the model fit, the χ^2 statistic along with the df, p-value and the χ^2 to df ratio are reported, as they are frequently used statistics to evaluate model fit.

The model fit turns out to be good for the X-Lex, the C-test and the reading test. For listening, the SRMR indicates a bad fit, while all other indices support the assumption of unidimensionality. Since the separate CFAs for the non-overlapping item sets strongly support unidimensionality, we decided to treat listening comprehension as a unidimensional construct despite the ambiguous results in the overall model.

For the joint model including all variables, most fit statistics could not be estimated when controlling for the sample structure due to data missing by test design and the relatively small number of classes. The SRMR supports the assumed factor structure. To provide more information on the complete model, fit statistics without control for the sample structure are additionally reported in Table 2; they also support the specified model. All other results, particularly statistical significance, are based on analyses controlling for the sample structure. Within all CFA models, all factor loadings were statistically significant, indicating that all indicators contribute to the measurement of their respective constructs.

Table 2. Model fit statistics resulting from the separate CFAs for each test and for the joint analysis of all tests.

Model	*n*	CFI	TLI	RMSEA	SRMR	χ^2	*df*	*p*	χ^2 / *df*
X-Lex	556	1.000	1.000	.008	.005	4.16	4	.385	1.04
C-test	558	.999	.999	.043	.007	4.11	2	.128	2.06
Listening	2112	.997	.996	.014	.266	97.45	70	.017	1.39
Listening set 1	1149	.997	.996	.019	.034	49.12	35	.057	1.40
Listening set 2	963	.997	.996	.019	.030	47.15	35	.082	1.35
Reading	1247	.959	.951	.029	.067	133.90	65	.000	2.06
All tests	559	–	–	–	.077	–	–	–	–
All tests*	559	.958	.954	.029	.077	877.94	600	.000	1.46

*Without controlling for the sample structure.

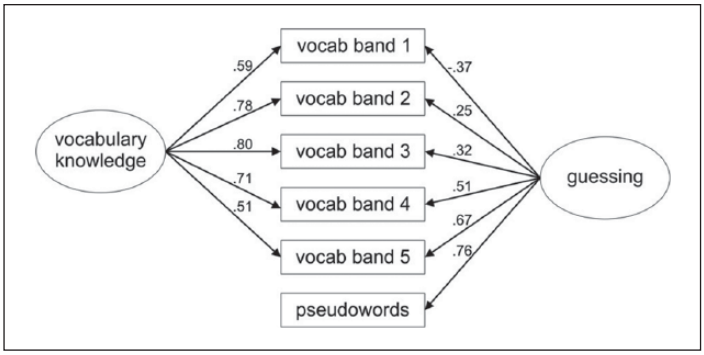

Figure 3. Two-dimensional CFA model for the X-Lex with completely standardized parameter estimates.

Since the model for the X-Lex is the only one more complex than a standard unidimensional model, the specified structure and the resulting model parameters are displayed in Figure 3.

Interestingly, the loadings of the vocabulary indicators on the guessing dimension increase with the vocabulary band. It seems that the less familiar the vocabulary items get, the more the score is affected by a tendency for guessing.

What is the relationship between the different measures gained from the tests for listening and reading comprehension, C-test and the Yes/No vocabulary test?

In order to examine research question 1 (i.e., the relationship between the four measures in our study), we conducted correlational analyses, both for the observed scores (Pearson's product–moment correlation) and at the latent level by means of SEM. Table 3 shows the observed-score correlations, Table 4 shows the latent correlations between the above detailed latent CFA dimensions.

Table 3. Correlations between the observed scores.

Observed scores	(1)	(2)	(3)	(4)
(1) X-Lex Hit Rate				
(2) X-Lex False-Alarm Rate	.327***			
(3) C-Test Total Score	.482***	–.155*		
(4) Listening Score (WLE)	.491***	–.097 n.s.	.762***	
(5) Reading Score (WLE)	.386***	–.158 **	.728***	.676***

Note: *$p \leq .050$; **$p \leq .010$; ***$p \leq .001$; n.s. $p > .050$; two-tailed tests.

Table 4. Latent correlations between the CFA dimensions.

Latent correlations	(1)	(2)	(3)	(4)
(1) X-Lex Vocabulary				
(2) X-Lex Guessing	+			
(3) C-Test	.638***	–.167 n.s.		
(4) Listening	.639***	–.140 n.s.	.863***	
(5) Reading	.624***	–.237**	.916***	.913***

Note: *$p \leq .050$; **$p \leq .010$; ***$p \leq .001$; n.s. $p > .050$; two-tailed tests.
+The correlation between the vocabulary factor and the pseudoword factor is fixed to zero in SEM.

In both approaches, the C-test correlates higher with the reading and listening measures than the X-Lex measures. For the observed scores, the highest correlations exist between the C-test and listening, followed by C-test and reading; the third highest correlation is found between listening and reading. For the latent correlations, the closest relationship is found between C-test and reading, followed by listening and reading, with C-test and listening coming at third place. X-Lex shows moderate correlations with all other measures. It does not appear to be closer related to the C-test than to the two receptive measures. It should be noted that the standard errors (SE) for all latent correlations are very small and the confidence intervals (CI) for even the highest latent correlations do not exceed 1.00 (e.g., for C-test with reading: SE = .028; 95% CI = .883 − .949). This means that despite the high correlations, the constructs assessed by the different tests can be empirically differentiated.

It is noteworthy that the hit rate (HR) correlates positively with the false-alarm rate (FAR), indicating that high HRs in the X-Lex are partially owing to guessing. Thus, controlling for the FAR appears necessary when using the X-Lex HR as predictor. It is interesting to note that the FAR is slightly negatively related to the C-test and reading, meaning that students with a tendency to guess in the X-Lex show a slightly poorer performance in the other tests.

Which of the two measures, C-test or Yes/No vocabulary test, is the more effective predictor for listening and reading comprehension for the observed scores?

Next, to examine research question 2 (i.e., the predictive power of the C-test and X-Lex measures for the listening and reading comprehension tests for the observed scores), we

Table 5. Standardized regression coefficients and explained variance in the regression models predicting listening and reading scores with the X-Lex scores and the C-test score.

Predictors	Dependent variable listening			Dependent variable reading		
	Model 1 X-Lex	Model 2 C-test	Model 3 both	Model 1 X-Lex	Model 2 C-test	Model 3 both
X-Lex hit rate	.59***	–	.19***	.49***	–	.09*
X-Lex false-alarm rate	–.29***	–	–.06 n.s.	–.32***	–	–.08*
C-test	–	.76***	.66***	–	.73***	.67***
R^2	.32***	.58***	.60***	.24***	.53***	.54***

Note: *$p \leq .050$; **$p \leq .010$; ***$p \leq .001$; n.s. $p > .050$; two-tailed tests.

conducted regression analyses. As outlined above, three models were analysed, each taking both listening and reading as dependent variables: (1) a model with the X-Lex HR and FAR as predictors, (2) a model with the C-test score as predictor, and (3) a model with the scores from both tests (i.e. X-Lex HR, X-Lex FAR and C-test score) as predictors. Table 5 displays the standardized regression coefficients and the explained variance from the observed score analyses.

Findings are very similar for listening and reading. Using the X-Lex alone to predict performance, the HR has a positive effect, while the FAR used as control variable is significantly negatively related to both dependent variables. The C-test score alone, however, as was to be expected from the correlation analyses above, can explain substantially more variance than the X-Lex scores. Both measures together can explain only a slightly higher amount of variance than the C-test alone, meaning that there is practically no explained variance unique to the X-Lex.

On a latent level, which of the two measures, C-test or Yes/No vocabulary test, is the more effective predictor for listening and reading comprehension?

To examine research question 2 on the latent level, we then conducted latent regression analyses using SEM. With respect to the dependent variables, the models were specified parallel to the observed score regression reported above. Again, three models were analysed, each using both the listening and reading dimensions as dependent latent variables. The first model contained only the X-Lex vocabulary and guessing factors as predictors, the second only the C-test factor, and the third all three dimensions. Table 6 shows standardized regression coefficients and explained variances.

With the X-Lex factors as predictors, the vocabulary dimension has a substantial effect on both dependent variables, while the guessing dimension has a small negative effect on reading, and no significant effect on listening. It is worth noting that despite the small or non-significant effects of the guessing factor, controlling for this factor has a substantial effect on the explanatory power of the X-Lex. If the FAR and hence the guessing factor are dropped from the X-Lex model, the explained variance drops to .28 for listening and .23 for reading.

Table 6. Standardized regression coefficients and explained variance for the prediction of listening and reading with the X-Lex factors and the C-test factor.

Latent predictors	Dependent latent variable listening			Dependent latent variable reading		
	Model 1 X-Lex	Model 2 C-test	Model 3 both	Model 1 X-Lex	Model 2 C-test	Model 3 both
X-Lex vocabulary	.64***	–	.15*	.62***	–	.09 n.s.
X-Lex guessing	–.14 n.s.	–	–.01 n.s.	–.24**	–	–.10**
C-Test	–	.86***	.76***	–	.92***	.85***
R^2	.43***	.75***	.76***	.45***	.84***	.85***

Note: *$p \leq .050$; **$p \leq .010$; ***$p \leq .001$; n.s. $p > .050$; two-tailed tests.

With respect to the comparison of X-Lex and C-test performance, the results from the latent variable regression confirm the above-reported observed-score regression results. The C-test has a markedly higher predictive power for listening and reading compared to the X-Lex. Again, there is hardly any variation in both dependent variables that can be explained by the X-Lex over and above the variance explained by the C-test alone.

Discussion and conclusion

In our study, we compared the performance of two instruments often reported as being suitable for screening and placement purposes, that is, the C-test and the X-Lex, a Yes/No vocabulary test (e.g., Alderson, 2005; Eckes & Grotjahn, 2006). While the predictive power of both measures for reading is widely reported in the literature, we also investigated their power to predict listening test scores and listening abilities. With regard to the question of which of the two formats is the better predictor of receptive skills, the literature does not report a direct comparison of the two predictive measures, which is why we investigated this aspect.

In answer to research question 1 (i.e., the relationships among the four measures in our study), both predictive measures show significant correlations with the reading and listening measures, broadly in line with what is reported in the literature (e.g., Alderson, 2005; Eckes & Grotjahn, 2006). The correlations are, however, not high enough to assume that any of our measures would load on the same dimension. While each instrument captures a distinct construct, the relationships between the constructs vary. When examining the relationship between C-test and X-Lex, we found moderate correlations for the observed scores and moderately high ones for the latent level. This, we would interpret, is an indicator that the two instruments measure distinct constructs.

We found that the C-test correlates higher than the X-Lex with our reading and listening comprehension measures, the X-Lex nevertheless showing substantial correlations. One possible explanation of the close relation between the C-test and the measures of receptive skills can be found in the nature of the C-test. It draws on a complex array of skills, many of which are also reported for reading processes (e.g., Eckes & Grotjahn, 2006; Sigott, 2004). Inferring from our results, these skills may be relevant for listening

processes, as well. The embedded, contextualized processing (Read, 2000; Read & Chapelle, 2001) necessary for solving the C-test items may be more closely related to the processing of spoken and written input than is the discrete and decontextualized processing associated with the X-Lex. Here, our findings differ from those reported in Qian (2008) who found only small differences in contextualized vs. decontextualized vocabulary measures.

With regard to research question 2 (which targeted the predictive power of the C-test and X-Lex for explaining variance in measures of receptive language skills), our findings are in line with research reporting the two formats to be good predictors of reading comprehension (e.g., Alderson & Huhta, 2005; Eckes, 2014; Meara, 2005b). Moreover, we found that they also can explain substantial variance in listening comprehension test scores, a result which chimes with Alderson's (2005) findings for a Yes/No vocabulary measure. When comparing the two predictive measures directly, our data suggest that the C-test outperforms the X-Lex both in traditional regression analyses and in SEM, with the latter showing substantially higher amounts of variance explained by both predictive variables since SEM controls for measurement errors. The superiority of the C-test over the X-Lex is equally clear in both observed-score regression and latent variable regression.

Comparing the predictive power of the two measures employed in our study across listening and reading comprehension, we found a slightly higher prediction for listening with the observed scores of both C-test and X-Lex, which may be attributed to the slightly higher reliability of the listening scores. When controlling for measurement error in SEM, the variance explained by the X-Lex on the latent variable level is almost identical for both listening and reading. For the C-test, the explained variance on the latent level is higher for reading, which may indicate the higher proximity of reading and the C-test on a construct level. The C-test's contextualized approach to language processing and the presentation of embedded items in their authentic textual context seems to account for a substantial amount of receptive language skills. We agree with Qian (2008) on the potentially positive washback effects of such contextualized test formats, since vocabulary phenomena which are presented in their natural context presumably provide a richer learning source than vocabulary items presented in isolation.

An additional interesting outcome of our analyses is that the simplest form of scoring a Yes/No vocabulary test, that is, calculating two separate scores for the hit rate (HR) and the false alarm rate (FAR), outperforms the four more complex methods suggested, for example, in Beeckmans et al. (2001) or Huibregtse et al. (2002). A further benefit of this approach is that the two resulting scores can be interpreted as vocabulary breadth and as guessing factor respectively, thus allowing correction for guessing, as suggested, for example, by Meara (2005b). A closer look at the correlation of the two scores for the X-Lex (i.e., HR and FAR) exhibits interesting patterns. The fact that they correlate positively with each other was to be expected. It is interesting to note that the more words a test taker ticks as known, the higher the FAR seems to become. Our data also indicate that the less familiar the vocabulary items get, the more the score is affected by a tendency for guessing. However, the fact that the FAR correlates negatively with the C-test and the receptive skills measures indicates that test takers with higher (receptive) language proficiency tend to make less use of guessing.

We have to concede that our study faced certain limitations. We had access to data from one educational context only, with a specific student sample limited to 14–16-year-olds in the higher secondary school tracks of Germany. Moreover, the range of test instruments was limited to two predictive measures and to tests of reading and listening comprehension. Hence, further research is needed for other contexts, test-taker groups, and test instruments. It would be particularly interesting to conduct a similar study with writing and speaking tests as dependent measures, in order to examine the predictive power of the C-test and the X-Lex for the productive skills. Here, it would also be worth examining measures of productive vocabulary knowledge and measures of vocabulary depth. Another limitation of our study is found in the paper-pencil administration of our test instruments. Thus, the predictive power of the C-test and X-Lex would need to be examined in computer-based administrations to allow insights into the potential of the formats in online environments, where they could be used as placement tests for computer-adaptive testing procedures.

To sum up, our findings indicate that the C-test seems to be the more powerful predictor for tests of receptive skills, outperforming the widely reported Yes/No vocabulary format. This has implications for the use of C-tests for placement or screening purposes. The C-test is reported to be a highly economical and robust instrument (e.g., Grotjahn, 2002), easy to develop, simple to administer, and reliable to score, which makes it an interesting format for test developers and administrators. The amount of explained variance by the C-test for both reading and listening, along with its reported qualities, indicates at least for our sample and instruments that the C-test is highly suitable as a screening or placement test with regard to predicting receptive skills.

Acknowledgements

We would like to express our gratitude to Paul Meara for kindly providing the items for the X-Lex and for his valuable feedback on scoring and analysis, as well as on an earlier draft of this manuscript.

We would like to thank the Institute for Educational Quality Improvement (IQB), Humboldt-University Berlin, and the Research Data Centre (FDZ) at the IQB, which provided the data set from the large-scale assessment study in 2008. We would like to thank the anonymous reviewers for their insightful comments and suggestions.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Alderson, J. C. (2005). The vocabulary size placement test. In *Diagnosing foreign language proficiency* (pp. 79–96). London: Continuum.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, *22*(3), 301–320.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535–556.

Asparouhov, T., & Muthén, B. (2005). *Multivariate statistical modeling with survey data*. Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference.

Author et al. (2007). **[to come]**

Author et al. (2008). **[to come]**

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No test: Some methodological issues in theory and practice. *Language Testing*, *18*(3), 235–274.

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

Coleman, J. A. (1996). *Studying languages: A survey of British and European students: The proficiency, background, attitudes and motivations of students of foreign languages in the United Kingdom and Europe.* (pp. 136–161). London: CILT.

Council of Europe. (2001). *A Common European Framework of Reference for Language Learning and Teaching*. Cambridge: Cambridge University Press.

Eckes, T. (2014). 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] (pp. 137–162). Frankfurt/Main: Lang.

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

Elder, C., & von Randow, J. (2008). Exploring the utility of a Web-based English language screening tool. *Language Assessment Quarterly*, *5*(3), 173–194.

Frey, A., & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35*(2), 89–94.

Grotjahn, R. (2002). Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis [Construction and use of C-test: a guide for practical implementation]. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* [The C-test. Theoretical foundations and practical applications]. Bochum: AKS, 211–225.

Harrington, M., & Carey, M. (2009). The online Yes/No test as a placement tool. *System*, *37*(4), 614−626.

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

Harsch, C., Pant, H. A., & Köller, O. (Eds.). (2010). *Calibrating standards-based assessment tasks for English as a first foreign language. Standard-setting procedures in Germany*. Münster: Waxmann.

Harsch, C., & Schröder, K. (2007). Textrekonstruktion: C-Test. In: B. Beck & K. Klieme (Eds.), *Sprachliche Kompetenzen: Konzepte und Messungen. DESI-Studie* (pp. 212–225). Weinheim: Beltz.

Harsch, C., & Schröder, K. (2008). Schülerkompetenzen im Englischen: Textrekonstruktion: C-Test. In: DESI-Konsortium (Ed.), *Ergebnisse der DESI-Studie* (pp. 149–156). Weinheim: Beltz.

Hastings, A. J. (2002). Error analysis of an English C-test: Evidence for integrated processing. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* [The C-test. Theoretical foundations and practical applications]. Bochum: AKS, 53–66.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a Yes–No vocabulary test: Correction for guessing and response style. *Language Testing*, *19*, 227–245.

Karimi, N. (2011). C-test and vocabulary knowledge. *Language Testing in Asia*, *1*, 7–38.
Kiefer, T., Robitzsch, A., & Wu, M. (2014). *TAM: Test Analysis Modules*. R package version 1.0–2. Retrieved May 9, 2014, from http://CRAN.R-project.org/package=TAM
Klein-Braley, C. (1985). A cloze-up on the C-test. *Language Testing*, *2*, 76–104.
Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*, 47–84.
Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing*, *1*, 134–146.
Köller, O., Knigge, M., & Tesch, B. (Eds.) (2010). *Sprachliche Kompetenzen im Ländervergleich* [Language proficiency in Germany: A comparison study]. Münster: Waxmann.
Kultusministerkonferenz der Länder (KMK). (2003). *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Mittleren Abschluss* [Educational Standards for the first foreign language English for the middle secondary school]. Darmstadt: Luchterhand.
Kunnan, A. J. (1998). An introduction to structural equation modelling for language assessment research. *Language Testing*, *15*, 295–332.
Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436.
Laufer, B., & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing*, *16*(1), 33–51.
MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.
Meara, P. M. (2005a). *X_Lex: the Swansea Vocabulary Levels Test. v2.05*. Swansea: Lognostics.
Meara, P. M. (2005b). *Lex vocabulary tests v2.0*. Swansea: University of Wales, Centre for Applied Language Studies.
Meara, P. M., & Jones, G. (1990). *The Eurocentres Vocabulary Size Test. 10KA*. Zurich: Eurocentres.
Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In: Daller, H., Milton, J. & Treffers-Daller, J. (Eds.). *Modelling and assessing vocabulary knowledge* (pp. 47–58). Cambridge: Cambridge University Press.
Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.
Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73–98.
Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide* (7th edn). Los Angeles, CA: Muthén & Muthén.
Nation, P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle & Heinle.
Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–81.
Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489–509.
Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: A structural equation modelling approach. *Language Testing*, *15*, 333–379.
Qian, D. D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly*, *5*(1), 1–19.
Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*, 28–52.
Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2013). Research timeline: Second language vocabulary assessment. *Language Teaching*, *46*(1), 41–52.

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

Reichert, M., Brunner, M., & Martin, R. (2014). Do test takers with different language backgrounds take the same C-test? The effect of native language on the validity of C-tests. In R. Grotjahn (Ed.), *Der C-test: Aktuelle Tendenzen. The C-test: Current trends* (pp. 109–136). Frankfurt/Main: Lang.

Rupp, A. A., Vock, M., Harsch, C., & Köller, O. (2008): *Developing standards-based assessment tasks for English as a first foreign language – context, processes and outcomes in Germany*. Münster: Waxmann.

Shiotus, T., & Weir, C. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99–128.

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

Sigott, G. (2006). How fluid is the C-Test construct? In R. Grotjahn (Ed.), *Der C-Test: Theorie, Empirie, Anwendungen* [The C-Test: Theory, empirical research, applications] (pp. 139–46). Frankfurt/Main: Lang.

Stemmer, B. (1991). *What's on a C-test taker's mind? Mental processes in C-test taking*. Bochum: Brockmeyer.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Yan, D., von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: Chapman & Hall/CRC.

Appendix A

Instructions for the X-Lex test (originally presented in German since it is the language of schooling; translation by author)

You will now work on a test which measures your vocabulary knowledge.
You will look at a list of English words. Decide for each word whether you know it or not.
If you know the meaning of the word, tick ☑ YES; if you are not sure or don't know the meaning, tick ☑ NO.
BE CAREFUL: Not all of the words in the list are 'real' English words; some of them do not exist.
If you tick a word as ☑ YES which does not exist, we will deduct one point!
The aim of this test is to measure the breadth of your vocabulary.
Here is one example:

that	☑	YES	☐	NO
darrock	☐	YES	☑	NO
lessen	☐	YES	☑	NO

You know the meaning of the word "that" – in this case, you tick YES.
You don't know the word "darrock" – in this case, you tick NO.
You are not sure about the word "lessen" – in this case, also tick NO.