Iranian Journal of Applied Linguistics (IJAL), Vol. 16, No. 2, September 2013, 107-135

On the Use of Offline Short Tests for Scoring and Classifying Purposes

Faeze Safari
Ph.D. Student of TEFL, Shiraz University, Shiraz, Iran

Alireza Ahmadi [a]
Assistant Professor of TEFL, Shiraz University, Shiraz, Iran

Received 2 June 2013; revised 24 July 2013; accepted 20 August 2013

Abstract

In response to the increasing interest in and need for a practical brief measure in language testing, this study explored the properties of an offline short-form test (OSF) versus a conventional lengthy test. From the total of 98 vocabulary items pooled from the Iranian National University Entrance Exams, 60 items were selected for the conventional test (CT). To build the OSF, we created an item bank by examining the item response theory (IRT) parameter estimates. Data for the IRT calibration included the responses of 774,258 examinees. Upon the results of the item calibration, 43 items with the highest discrimination power and minimal guessing values from different levels of ability were selected for the item bank. Then, using the responses of 253 EFL learners, we compared the measurement properties of the OSF scores with those of the CT scores in terms of the score precision, score comparability, and consistency of classification decisions. The results revealed that although the OSF generally did not achieve the same level of measurement precision as the CT, it still achieved a desired level of precision while

[a] *Email address*: arahmadi@shirazu.ac.ir
Corresponding address: 7194684795, Department of Foreign Languages and Linguistics, Faculty of Literature and Humanities, Shiraz University, Shiraz, Iran

lessening the negative effects of a lengthy test. The results also signified an excellent degree of correspondence between OSF and CT scores and classification results. In all, findings suggest that OSF can stand as a reasonable alternative for a longer test, especially when conditions dictate that a very short test be used.

Keywords: Offline short form; Item response theory; Item parameter; Conventional test

Introduction

The desire for a shorter form of a test has been preceded by numerous efforts in the past decades, seemingly beginning when Doll (1917) first questioned if it was necessary to make use of all the Binet-Simon questions to estimate intelligence. Since then, test reduction efforts have appeared to be of growing interest for researchers of various domains. The term short form is reserved for those approaches in which a test has been reduced in length from an original full-length test. The widespread adoption of short forms highlights the point of practicality in administrating a test. Unfortunately, the very feature of most conventional tests renders them less feasible. On average, most conventional tests, particularly those developed for measuring language proficiency, are too long which makes them less practical at least in time sensitive contexts. Petway (2010) noted that a measure that is too long may lead to unwanted forms of bias and threaten the clarity of the results as the examinees may experience fatigue or lose interest while responding a measure. Concerns such as these support efforts to develop short form tests which serve to reduce test length, abbreviate the administration process, and consequently lessen the burden placed on examinees. The added benefit of short forms articulated by Petway (2010, p. 53) is that "it removes some of the extraneous noise gained from items that do not share as much information with the total score". In test reduction efforts, therefore, the focal issue is the indication of items which are functioning best as trait indicators and items which are not performing well psychometrically and thus should be considered for elimination. By and large, the task of short forms is to conserve time by removing items from the full test that do little to add to the measurement while sacrificing as little as possible the measurement precision. Indeed, this may lead to greater efficiency in measuring the examinees and obtaining scores (Haley, Coster, Andres, Kosinski, & Ni, 2004).

A short assessment can be beneficial where most of language tests then and now are quite lengthy, time-consuming, cumbersome, and unattractive which in turn,

would have detrimental effects on the outcomes. A good language measurement tool must be brief enough to be adopted in practice. These concerns were echoed and reaffirmed by Giouroglou and Economides (2004) who argued that "a test needs to be practical and economic. Maximum quality with less effort and within less time is preferable. Economy in time and item selection can result in increased test production and higher scores from the part of examinees" (p. 750). The need for shorter, psychometrically sound tests in language testing therefore appears high. Overall, if we select a powerful methodology for short-form development and build careful short forms, then, we can find scientifically supportable and hence valuable shortened tests (Smith, McCarthy, & Anderson, 2000).

Short-form Tests: A Methodological Review

Whereas the literature presents extensive guidance and discussions on understanding and developing a full-length test, issues related to a shortened version of a full-length test are rarely discussed to the same degree. With respect to language testing and specifically testing English as a foreign language, little or no mention is given to such concerns in reducing the length of a test. Nevertheless, one can organize the available short form tests, which have basically been developed for medical and clinical studies, roughly into two categories; one that uses classical test theory (CTT) in test construction procedures: CTT-based short tests and the other that employs IRT techniques in test construction: IRT-based short tests.

The construction of short tests using CTT involves the selection of the items according to classical indices. Different item selection strategies under this framework are available. Among them, the most common ones are item characteristics approach, correlational approach, and factor analytic approach. Despite their prevalence, the application of CTT item selection methods brings about some limitations on the test. A limitation of all these approaches is that the scores on the shortened tests will not be directly comparable with the scores from the full tests, because they are not on the same scale (Khan, 2010). Moreover, Hambleton, Swaminathan, and Rogers (1991) detailed problems of using the classical item characteristics approach. First, "classical indices are not invariant over populations that differ in ability" (p. 99). This means that if the group which is used to determine item characteristics is different from the target population for whom the test is to be used, "the item indices obtained will not be appropriate for the intended population" (p. 99). In other words, the success of such an approach

depends on the extent to which the group and the intended population match. Second, "the contribution of an item to the reliability of the test does not depend on the characteristics of the item alone, but also on the relationship between the item and the other items in the test" (p. 100). Hence, it is not possible to determine the sole contribution of a single item to the reliability of a test and its conceptual converse, standard error of measurement. Accordingly, CTT does not permit us to select items to build a test with a prespecified desired measurement precision by adding or removing a set of items, even if a well-constructed item bank is available to choose items from.

With respect to the other two approaches, namely correlational approach and factor analytic approach, Stanton, Sinar, Balzer, and Smith (2002) explained that such approaches consider internal consistency maximization in making short tests. They argued that selecting items to maximize internal consistency would result in excluding all the items but those highly similar in content and thus including a set of items which are highly redundant in appearance, narrow in content, and potentially low in validity. Another potential negative consequence argued by Weiss (2004) is that, in order to maximize internal consistency reliability, typically items which are suitable for the average test takers in the group (i.e., items with average item difficulty indices) are selected. In so doing, the items could be regarded as the best measurement of the average examinees, too difficult for less proficient examinees, and too easy for more proficient ones. The choice, in these methods, therefore is almost always in favor of the average examinees.

IRT provides a more powerful method of item selection in short tests than does CTT (Drake, 2011). Here, item parameters are invariant, overcoming the limitations of CTT indices reviewed above. Under the IRT framework, test reduction is very often done by online mode of testing. An online short test is a dynamic short-form test where item selection and ability estimation are ongoing processes. Such tests are almost always carried out by computers and hence they are known as computerized adaptive tests (CATs). The aim of such tests is to reduce the length of a test and increase the efficiency of the test by providing for each examinee a set of informative items that measure the individual on the trait effectively (Weiss, 2004).

Despite the strength of CAT performance, there are still many educational communities which are slow to move into CAT due to the quite demanding process

of CAT development, both technically and financially. It takes a high level of knowledge and expertise, time, and money to develop an item bank and launch a CAT project (Dunkel, 1997). The lack of computer availability for the test administration may also impose some restrictions on the applicability of a CAT system. Whereas with a big room, a large number of individuals can take a paper and pencil (P&P) test, with a computer lab, seating capacity is more limited and a mass single administration may not be possible, resulting in the administration of the CAT two or more times. This may cause a group of students to be exposed to the same items which, in turn, makes it possible that the students could respond an item correctly based on their previous knowledge of the item rather than based on the ability that was intended to be measured (Johnson, 2006). This makes test/item security an issue of concern for CAT.

The difficulty of reviewing test form quality is another possible limitation of a CAT project. Zhang (2006) points out that an item bank as a whole can be examined and checked before its activation in a CAT project, but every examinee's test form cannot be reviewed in advance because test forms do not come prepackaged but rather are individually designed in an on-going process. Thus, during the CAT administration, there is no opportunity for intervention. It has also been discussed that CAT is not suited for extended response type of items like essays because they cannot be scored online.

Thus, the need for an approximation or alternative to CAT that could be more straightforward and low-tech arose and attempts were made to develop a short test that mimic the CAT procedure in an offline mode, i.e. without computer and in a traditional P&P environment. This test is known as a static IRT-based short form, or in Padaki and Natarajan's (2009) terms an offline test. An offline short form test (OSF) as an alternative short test design that purportedly addresses CAT's nonpsychometric shortcomings is primarily used to achieve measurement efficiency, especially in the absence of computer technology.

The basic notion of an OSF is to mimic what a CAT would do in an offline and fixed mode. The idea of offline mode of short form has been around probably as far as the IRT framework is concerned. In recent years, however, the findings of Reise and Henson's (2000) study coupled with concerns about the practical shortcomings of CAT have attracted interest in this alternative form of testing. Reise and Henson (2000) found that the four best items, i.e. the most discriminating items, were

selected most often when the CAT version of the Revised Neuroticism-Extroversion-Openness Personality Inventory (NEO PI-R) was administered. This finding suggested that a fixed short form of these four items would have performed as equally well as the CAT version of NEO PI-R.

By making slight modifications in the process of CAT, an OSF is implemented in a P&P environment which makes it more practical (Padaki & Natarajan, 2009). Since here the test is offline, items are not selected as the test proceeds but they are selected and prepackaged prior to the test administration, though the item selection rules are the same as those in CAT. Also, the examinee's score, in the OSF system, is not estimated during the test after responding to every single item but it is estimated after responding to all the test items, though the rules for obtaining the scores are the same as CAT score estimation rules.

Methods applied for item selection in OSFs, like CAT, are primarily based on IRT-based item parameters, i.e. item difficulty (*b*), item discrimination (*a*), and guessing (*c*), and their corresponding curves, namely item characteristic curve (ICC) and item information function curve. A review of the literature shows clearly that the item *a*-parameter is the major factor influencing item exposure in a CAT system (e.g., Chang & Ying, 1999). *a* parameter provides the amount of information an item would yield in the given *b* value. Selecting items with maximum information in CAT leads to a substantial gain in efficiency (Chang & Ying, 1999). Hence, Hol, Vorst, and Mellenbergh (2007) assert that it is possible that OSFs created on the basis of *a*-parameter values perform equally well compared to CATs. Of course, in addition to the *a* values, it is important to balance the range of *b* values to produce a wide distribution of difficulty levels across the ability continuum; thus ensuring a good match between the items and different ability levels (Chang, Qian, & Ying, 2001). With respect to the number of test items, it should be mentioned that overall, an OSF requires a bit larger number of items than a CAT to achieve comparable precision because in general, an OSF attempts to optimize larger areas of ability at once whereas CAT maximizes information about every single examinee separately (Choi, Reise, Pilkonis, Hays, & Cella, 2010). Thus, the main concern in an OSF is to develop a shortened version of a test brief and simple enough to be practical, yet broad enough to comprehensively measure a wide range of ability and cover all the content areas included in the full-length test (Kosinski et al., 2003).

If one were to put the degree of score precision of the shortened tests on a continuum, then traditional CTT-based short forms would be at one extreme, where test reduction techniques are some form of internal consistency maximization, resulting in the selection of a set of items which are highly redundant in appearance, narrow in content, and potentially low in validity (Stanton et al., 2002) and where the tests are almost always in favor of the average examinees with no point of adaptation to different levels of ability. On the other extreme would be the CATs where, in Keng's (2008) terms, "adaptation occurs after every item and each examinee could potentially receive a completely different set of items" (p. 46). Offline short tests would then take the middle ground between these two extremes. An OSF does not take its points of adaptation at the individual levels, but instead it tries to optimize to a larger group at once. OSFs rely on a fixed set of questions that may not possibly be the best set for all the examinees, but they instead try to achieve their best precision by spreading high quality questions over a relatively broad range to cover higher and lower levels of ability as well as the average level (Jette, 2003).

Given the current interest in OSF in medical and clinical-based studies, it is noteworthy that it has been virtually ignored in the field of foreign language testing where, on the one hand, the use of markedly lengthy tests is a big issue for measurement and research settings and on the other hand the use of CAT to make the test shortened is apparently unfeasible in many contexts. Therefore, it is fair to ask, why are not more applications in the realm of language testing and specifically foreign language testing if IRT-based short forms are so advantageous and so successfully used in other settings? Consequently, our chief purpose in the present study was to explore the properties of an OSF in comparison with a conventional lengthy test (CT) when the tests are to measure the vocabulary knowledge of EFL learners. The two testing systems were compared in terms of: (1) score precision, including the standard error (SE) of ability (θ) estimates, the bias of the ability estimates (differences of OSF ability estimates from the CT ability estimates), and ceiling and floor effects, (2) score comparability, the correlation between OSF and CT ability estimates, and (3) quality of classification decisions, in terms of classifying examinees into two groups (masters/nonmasters) and three groups (below basic, basic, and proficient). Toward that end, the following research questions were put forward:

1. How precise are the OSF scores in comparison with the CT scores?
2. How well do the OSF scores represent the CT scores?

3. How accurate are the results of OSF classification decisions in comparison with those of the CT?

Method

Instrumentation

Creating item bank. To start the study, we had to build an OSF out of an item bank. Development of an item bank therefore was the first step. This was done in four phases. The first course of action was item development. For this step, the study pooled the items from the Special English subtests of the Iranian National University Entrance Exam (INUEE) from three consecutive school years (2003, 2004, and 2005). The aim of the subtest is to measure candidates' general English language proficiency to screen the best for admission to English majors. Each subtest consists of 70 dichotomously-scored MC items with four alternatives in different areas including: structure, vocabulary, word order, language function, cloze test, and reading comprehension. As the focus of this study was on estimating vocabulary knowledge of the examinees, all of the analyses were directly performed on the items testing vocabulary, i.e. a total of 98 items in the three years.

The second phase in the item bank construction was collecting data for IRT calibration. Data was obtained from the archive of the National Organization for Educational Measurement of Iran. The dataset included the responses of 774,258 examinees who sat for the Special English subtests mentioned above. In detail, the data included a general population of 270,201 examinees taking the 2003 test, 284,403 cases taking the 2004 test, and 219,654 cases sitting for the 2005 test. Calibration data overlapped somewhat by including a number of examinees who sat for at least two of the tests. This is common-examinees linking which makes the items of the three tests be calibrated on the same scale (Davey, Oshima, & Lee, 1996).

The next phase was item calibration and test of model fit. Using BILOG-MG, (Zimowski, Muraki, Mislevy, & Bock, 2003), IRT parameter estimates were computed by testing the statistical fit of one-, two-, and three-parameter logistic models for each item. The main purpose of calibration at this stage was to identify poorly performing and misfitting items so as to remove them from the item set. The results indicated the three-parameter logistic model as the best fitting. Also, the test information functions (TIFs) for the three years exams from which the 98 items were pooled revealed that an abundance of the high quality items was covering moderate to difficult levels of proficiency and very few informative items were

extending to the lower end of the continuum. That is, the items provided the most information (and therefore the highest precision of measurement) for the people at the higher ability. And, the exams exhibited the most prominent loss of information for θ levels lower than – 0.5.

In the end, for the purpose of developing the item bank, we retained a set of 43 items with nearly high discrimination values and with item difficulties represented at various levels across the ability continuum of the target sample.

Building CT. In the present study, the conventional test (CT) was considered as a benchmark to evaluate the success of the OSF in estimating the vocabulary knowledge of the examinees. Typically, in literature, a full test which consists of all the available items is used as the criterion for short form performance (e.g. Haley, Coster, et al., 2004; Wang, 2009). However, in this project, considering the roughly large size of the full test and in order to increase the response rate on the part of the participants, it was supposed desirable to develop a shorter form to maximize time saving. Therefore, we drew 60 items from the full test for inclusion in the CT. Test reliability estimated through Cronbach's alpha turned out to be sufficiently high (0.92).

Building OSF. Once the item parameters were estimated and the final item bank was specified, we tried to develop the OSF. Note that if an OSF does not have adequate breadth over the θ scale, it will possibly lead to ceiling and/or floor effects and subsequently, failure to capture all the relevant ability levels. Hence, items of maximum information across the θ scale were selected from the bank for inclusion in the OSF.

It was expected that an OSF with 12 items would constitute an abridged test that would still exhibit acceptable levels of measurement properties. To ensure that there were adequate items available at each level of ability, the item bank was divided into 12 blocks based on *b* values. The blocks as well as the items within each block were arranged in ascending order while keeping items with high *a* values and low *c* values within each block. Therefore, the first block included items with the lowest *b* values, and the 12th block included items with the highest *b* values. The adjacent blocks overlapped partially in terms of the difficulty of the items. Each item was selected from one block for inclusion in the OSF. Beyond considering item properties in the selection of the items for an OSF, it was

important to take steps to ensure that the content of the long test was adequately represented in the OSF (von Baeyer, Chambers, & Eakins, 2011). Hence, while item selection was carried out, the content of the items was reviewed to ensure that the shortened test conserved content coverage.

We decided to create 30 short forms of the OSF from the available items in the bank. Using 30 different test forms would allow us to estimate and compare scores of the examinees who answered different sets of items in order to have a better understanding of the overall measurement properties of the OSF with different combinations of the items. Overall, the end result of this step of the study was the creation of 30 short forms from the available items in the bank so that each short form consisted of 12 high quality items distributed uniformly across the ability scale and arranged in a sequence of increasingly difficult items.

Participants and Data Collection Procedures

The sample used for the study involved a total of 332 examinees consisting of 73 upper intermediate EFL learners from four different English language teaching institutes and 259 English-major freshmen students from eight universities across the country, Iran. Since the tests used in the study pooled items from the English subtests of the INUEE, participants were selected in a way to roughly match this level of vocabulary knowledge. The participants were not needed to be perfectly homogenous, bearing in mind that all of the participants took two tests (OSF and CT) and the purpose of the study was to see how the performances on the two tests married up. As such, only an overall match between the participants' levels of vocabulary knowledge and the test difficulty level was required. All the examinees volunteered for the participation in the data collection process.

All the participants received two tests, CT and one short form, but the tests were administered to the examinees in a random pattern, with approximately half of the examinees taking the CT first and half taking one of the short forms first, followed by the other test. Administration was done in this manner to prevent order effect on test performance and to counterbalance fatigue effects. After removing the invalid answer sheets, 253 examinee responses remained for data analysis. Invalid responses included those with great mismatch between the responses to common items in the two tests (the probability of answering by chance), those with incomplete answer sheets, those with unusual response patterns, e.g. a repeated pattern for every four items, and finally those which were completed in an

extremely short time. Also, we had to discard short form 19 from the data analysis since accidently few answer sheets remained for this short form after removing the invalid answer sheets.

Data Analysis

Analyses were conducted in two parts. First, item calibration was conducted to determine item characteristics in order to choose the best items for the item bank. The procedure was explained in detail in *Creating item bank* Section.

The second part of the analysis investigated the psychometric properties of the OSF using the response data of examinees with valid answer sheets, i.e. a total of 253 answer sheets. We compared the estimates of the vocabulary knowledge score of these 253 examinees in the CT and OSF in terms of their relative *precision*, *score comparability*, and *classification consistency*. In all the steps, the CT performances served as the basis against which OSF scores were compared and evaluated.

Though the scoring system for a conventional test is typically within the framework of number-correct scoring system, in the present study, the score estimates were based on IRT scoring for both formats of the tests so as to put all the scores on the same scale and to facilitate the comparison of the scores and classification decisions. An additional advantage of IRT-based scoring is that standard error is calculated for every individual separately. Among the IRT scoring approaches, we preferred to use the Expected A Posteriori (EAP) approach which, unlike maximum likelihood estimation (MLE), could estimate the scores for examinees who answer all the items correctly or all the items incorrectly and produce more precise ability estimates (Wang, 2009). Obtaining examinee parameters (i.e. scoring examinees) was done through BILOG program using 1-PL model, an IRT model which commonly provides more stable estimates with the relatively small sample sizes (Haley et al., 2004). It should be noted that for the scoring phase, we added a command to the BILOG syntax to import the item parameters which were calculated during the first phase of the analysis into the program in order to use these known item parameters during the score estimation procedure.

To evaluate how precise the OSF scores are, three evaluative criteria were used: standard error (SE) of estimation, bias, and ceiling and floor effects. Bias was

defined as OSF score minus CT score. The values of bias indicated how far the OSF scores were from the CT scores (Cook et al., 2008; Wang, 2009). The positive bias values indicated the ability estimation process in OSF had a tendency to overestimate the θ and the negative values were a sign of underestimation of the θ values. Ceiling and floor effects were defined as the percentage of examinees who answered all or none of the items correctly.

To go beyond the overall analyses of the entire range of θ estimate, we estimated SE and bias for three intervals along the θ continuum: $\theta \leq -1$, $-1 < \theta \leq 1$, $\theta > 1$. This allowed us to have a better understanding of the OSF performance in different levels of the ability.

Score comparability was evaluated in terms of the Pearson product-moment correlation between the ability estimates from the CT and OSF. This variable indicates how well the OSF scores are similar to the CT scores. To ensure that the correlation was not inflated by the common items in the two tests, we removed the common items from the CT and then estimated the correlation.

At last, to evaluate how well the OSF classified examinees, two types of classification were made:

1. two-level classification in which the examinees with $\theta < 0$ were classified as nonmasters and those with $\theta \geq 0$ as masters, and
2. three-level classification where the vocabulary knowledge of examinees with $\theta \leq -1$, $-1 < \theta \leq 1$, $\theta > 1$ was considered 'below basic', 'basic', and 'proficient', respectively. The labels were taken from Zhang (2010).

Classification accuracy was measured by comparing the classification results of the OSF with the results of the CT to see how consistent the results of the OSF were. To evaluate the classification consistency, two evaluative criteria were used: the agreement coefficient (ρ) and Cohen's Kappa coefficient (κ).

Results

Comparison of OSF and CT

Score precision. The goal here was to know whether substantial reduction in respondent burdens is possible while maintaining acceptable standards of score precision. Precision of the OSF θ estimates was determined based on the SE, bias

of the estimates, and floor and ceiling effects in comparison to those of the CT. The results are presented in Table 1. As depicted, the mean of the scores for the OSF was 1.04 and for the CT was 0.85, indicating that the examinees achieved, on the whole, slightly higher scores on the OSF by an effect size of 0.25, a small effect size. With respect to the errors of the estimations, Table 1 shows that the SE for the OSF scores is 0.40 and for the CT scores is 0.20. As such, the conventional test estimated the scores with more precision than the OSF. In terms of bias, the distribution of the difference scores showed that there were more positive values than negative values, indicating that more examinees scored higher on the OSF than on the CT. This is reflected in the mean of the difference scores, reported in the table, where the bias is + 0.24. In greater detail, almost 68% of the examinees scored slightly higher on the OSF than on the CT, 23% of the examinees scored a bit higher on the CT than on the OSF and 9% of the examinees received the same scores on both tests. The person ability distribution also demonstrated no obvious floor and ceiling effects in the tests except a slight ceiling effect for the OSF where two individuals (0.8%) answered all the items correctly.

Table 1
Descriptive Statistics for Score Estimates, SE, Bias, and Ceiling and Floor Effects in CT and OSF (n=253)

Tests	Score Estimates		SE		Bias		Ceiling	Floor
	M	*SD*	*M*	*SD*	*M*	*SD*	(%)	(%)
CT	.85	.76	.20	.14	–	–	0	0
OSF	1.04	.77	.40	.05	+.24	.50	.8	0

For more detailed comparison of the tests along different ranges of the scores, analyses of the SE and bias are shown in Table 2 for three intervals along the θ continuum: $\theta \leq -1$, $-1 < \theta \leq 1$, $\theta > 1$.

Table 2
SE and Bias in CT and OSF across 3 Score Levels

Tests	SE			Bias		
	Low Score Range $\theta \leq -1$ (n=9)	Mid Score Range $-1 < \theta \leq 1$ (n=128)	High Score Range $\theta > 1$ (n=116)	Low Score Range $\theta \leq -1$ (n=9)	Mid Score Range $-1 < \theta \leq 1$ (n=128)	High Score Range $\theta > 1$ (n=116)
CT	.22	.18	.20	–	–	–
OSF	.42	.39	.41	-.50	-.34	-.13

The table shows that the SE of both tests increased as the scores moved away from the center of the ability distribution. That is, both CT and OSF estimated the scores with *slightly* more precision for individuals scoring below +1 and above -1 than for individuals at the two tails of the ability continuum. In other words, the most precisely measured examinees were those with θ values around zero while examinees with θ values at the two extremes of the scale were measured with *slightly* less precision. A review of the bias of the test across three intervals indicated that the smallest bias was at θ levels above 1 and the largest bias was at ability levels below -1. In other words, scores were the most biased for the more extreme negative θ values and the least biased for the more extreme positive θ values.

Score comparability. Similarities between the scores of the two tests were examined by computing the Pearson correlation coefficient to assess the extent to which OSF scores were consistent and comparable with scores from the CT. Pearson correlation coefficient was calculated for each set of the scores from the OSF-generated short form and the corresponding examinees' scores in the CT. The results are given in Table 3. As the Table indicates, all the correlations between the CT and short forms were significant and substantial except for those relating to short forms 7, 10, and 25. Apart from these three short forms, the correlations between the CT and short forms were quite high, ranging from 0.76 for short form 2 to 0.99 for short form 21, with an average of 0.86. In terms of the variance, OSF, on average, accounted for 74% of the variance in the CT which according to Cohen (1992), this amount of variance represents a very large effect size. This suggests that performance on the OSF is associated strongly with performance on the CT.

Table 3
Correlation Coefficients between CT and OSF Scores

Short Form NO.	Correlation Coefficient (*r*)	*P*	Short Form NO.	Correlation Coefficient (*r*)	*P*
Short form 1 (n=9)	**.901**	.001	Short form 16 (n=8)	**.760**	.049
Short form 2 (n=8)	**.762**	.040	Short form 17 (n=8)	**.883**	.003
Short form 3 (n=8)	**.822**	.012	Short form 18 (n=9)	**.853**	.004
Short form 4 (n=8)	**.884**	.004	Short form 20 (n=9)	**.762**	.027
Short form 5 (n=9)	**.795**	.010	Short form 21 (n=8)	**.990**	.001
Short form 6 (n=10)	**.863**	.001	Short form 22 (n=8)	**.981**	.001
Short form 7 (n=10)	.608	.062	Short form 23 (n=8)	**.879**	.004
Short form 8 (n=8)	**.885**	.008	Short form 24 (n=10)	**.913**	.000
Short form 9 (n=10)	**.851**	.002	Short form 25 (n=8)	.501	.255
Short form 10 (n=10)	.532	.351	Short form 26 (n=8)	**.773**	.046
Short form 11 (n=10)	**.813**	.004	Short form 27 (n=8)	**.978**	.001
Short form 12 (n=9)	**.833**	.038	Short form 28 (n=10)	**.771**	.009
Short form 13 (n=8)	**.928**	.007	Short form 29 (n=8)	**.859**	.006
Short form 14 (n=9)	**.945**	.000	Short form 30 (n=9)	**.850**	.004
Short form 15 (n=8)	**.801**	.041			

Notes. n: the number of participants for each short form; The sample size varied among short forms because for each short form, a number of invalid answer sheets (from 1 to 4) were set aside; Short form 19 was totally set aside from the data analysis procedure since a few answer sheets remained for this short form after removing the invalid answer sheets; *p* = level of significance; Significant *r* values are bold.

Recall that, in order to ensure that the correlations were not inflated by the common items in the two tests, we first removed the common items from the CT and then estimated the correlation. According to Smith et al. (2000), in so doing,

we may underestimate the overlap. That is, the real amounts of correlation between the two tests would probably be higher than those reported here.

Classification consistency. The classification results of the OSF were compared with those of the CT in order to obtain the extent of agreement between the two tests in classifying the test takers into two categories of master and nonmaster, and three categories of below basic, basic, and proficient.

Results from classifying examinees as mater/nonmaster according to the OSF and CT are displayed in Table 4. The table indicates that the CT classified 38 students as nonmasters and 215 cases as masters while the OSF classified 22 students as nonmasters and 231 ones as masters. In terms of classification consistency, 229 (= 18 + 211) of the sample (n = 253) are classified in the same way in the two tests. Out of all the students in the master category (n = 215, based on the results of the CT), the OSF correctly classified 211 students (98.2 %) as masters. And, the proportion of correctly identified nonmasters was 47.5% (18 students out of 38 students in the nonmaster category). Of the misclassified cases (n = 20 + 4), most were nonmasters incorrectly classified as masters in the OSF; that is, most of the errors were of false positive type. On the whole, the classification of 91% of the test takers remained unchanged across the two tests and the agreement coefficient (ρ) reached 0.91, indicating a very high agreement between the two tests in classifying the students into two categories. However, the Cohen's Kappa coefficient (κ) for this concordance was 0.55. The Kappa coefficient estimated a substantially lower reliability since it adjusts for the expected chance agreement (Ary, Jacobs, Sorensen, & Razavieh, 2010).

Table 4
Master/Nonmaster Classification in OSF and CT

		CT		
		Nonmaster	Master	Total
OSF	Nonmaster	18	4	22
	Master	20	211	231
	Total	38	215	253

Note. Cell values indicate the number of examinees in each category.

Table 5 presents the results of the OSF and the CT in classifying the examinees into three categories. Here, the proportions of correctly identified cases were 66.67% (6 out of 9 students) for the below basic level, 84.38% (108 out of 128 students) for the basic level, and 92.24% (107 out of 116 students) for the proficient level. These results suggest that the OSF had substantially high accuracy in the classification of basic and proficient students but lower accuracy in correctly classifying examinees as below basic. Of the misclassification cases, all were misclassified into an adjacent category and none of the examinees was misclassified into a far distant category (e.g. a student in the Below Basic category being classified as Proficient). Overall, 221 students (87%) of the sample were correctly classified which yielded an agreement coefficient (ρ) of 0.87 and a Kappa coefficient (κ) of 0.75. Here, the agreement coefficient was slightly lower than that of the two-level classification, but surprisingly, the Kappa coefficient was considerably higher.

Table 5
Below Basic, Basic, and Proficient Classification in OSF and CT

		CT			
		Below Basic	Basic	Proficient	Total
OSF	Below Basic	6	0	0	6
	Basic	2	108	9	119
	Proficient	1	20	107	128
	Total	9	128	116	253

Note. Cell values indicate the number of examinees in each category.

The lower value of κ for the two-level classification is due to the great reliance of Kappa coefficient on the number of codes, i.e. the number of categories in the classification. Bakeman, Quera, McArthur, and Robinson (1997) asserted that "values for Kappa are lower when codes are fewer" (p. 357). Due to this limitation of the Kappa statistic, we relied on the results of the agreement coefficient (ρ) in the current study which indicated substantial agreement between the two sets of the tests in classifying examinees into two and three groups.

Discussion

The focus of the study was on examining the psychometric quality of the offline short form testing in comparison to the conventional testing. Specifically, there were three main areas of interest to examine: quality of person θ estimates,

comparability of θ estimates, and quality of the classification decisions. In what follows, we discuss the results obtained in light of the focused areas.

Quality of θ estimates: How precise are the OSF scores in comparison with the CT scores?

One of the critical questions was to what extent the OSF scores were precise. The analysis of the means of the two tests and bias of the score estimates revealed that examinees' CT scores were lower than OSF scores. That is, the examinees achieved slightly higher scores on the OSF than on the CT. From these results a tendency was seen for the OSF to be somewhat easier for the test takers. This pattern of scores is apparently the same as those found in Choi et al. (2010) and Wang (2009) but with a slightly smaller difference between the two test scores which probably was due to the smaller length of their conventional tests (28 and 39 items, respectively). This finding may in part be explained by the fact that probably the 60-item CT used in this study was long and demanding on the part of the test takers and consequently some examinees might have experienced fatigue and lost interest as they continued to answer the conventional test; this is probably true as the researchers observed that the test was seen by some students as too daunting to even start. Also, since longer tests take more time to answer, they often tend to have more missing data and have higher rates of guessing than short tests. It would be quite reasonable to expect the rate of the wrong answers, missing answers, and/or guessing to be higher for a lengthy test, like the CT used in the present study, specifically when the test is for research purposes rather than scoring purposes which makes some test takers be less responsive; evidence from this may also come from the fact that about 25% of the answer sheets were faulty and removed from further analysis. Cheating could have been another possible factor as well. In the CT, the examinees received the same questions, so there was a chance for cheating, especially for the examinees who had lost their interest during the test administration and were therefore encouraged to cheat, whereas in the OSF the examinees' chance of cheating was nearly reduced to zero because they received different short forms. All these undesirable factors could have affected the test performance, threatened the scores, and consequently resulted in obtaining lower scores on the CT. On the other hand, decreased time demands on the examinees in the OSF might have increased the probability of answering the test with much more attention, resulting in fewer wrong or missing answers, and guessing. The researchers also observed that most of the examinees answered the OSF questions with enthusiasm when they found the test is so short. This issue is reaffirmed by

Giouroglou and Economides (2004) who maintain that "economy in time and item selection can result in increased test production and *higher scores* [emphasis added] from the part of examinees" (p. 750). This may suggest that there is the possibility that OSF provides potentially more valid inferences about the individuals' performances than a lengthy CT (if we can assume that answering with much more attention resulting in more valid interpretations).

Another interesting result obtained from the bias of the OSF scores is related to the pattern emerged for the bias along the θ continuum. Bias values decreased in magnitude as θ estimates increased. In comparison to the CT, OSF showed the biggest gap in measurement at the low score range ($\theta \leq -1$); but, while approaching the positive end of the continuum, OSF showed the least biased scores. In loose terms, the difference between the scores obtained from the CT and OSF was the biggest for the examinees in the lower end of the θ scale ($\theta \leq -1$); the difference then decreased for those in the middle of the scale ($-1 < \theta \leq 1$) and reached its lowest value for those in the higher end of the scale ($\theta > 1$). Remember that the extent to which the scores of the two tests are similar indicates how well the OSF functions in estimating the examinees' scores. The pattern emerged for the bias in the present study runs somehow counter to those of Hol et al., (2007) and Wang (2009) who found that a short form like OSF performs better for midrange scores in terms of the score bias as they observed that the difference between the short test and the conventional test was the least for the examinees in the middle of the scale rather than in the higher end of the scale. The difference in the results may in part be explained by the quality of the items used in the present study where the total item information curves had a quite peaked information function at the positive part of the scale. That is, the exams from which the items were drawn had the most informative items for those in the middle to higher end of the scale and few items providing high information for test takers with low ability were available. Then, the influence of item characteristics on the short forms' scores became apparent in the pattern of bias of θ estimates where the bias became increasingly lower towards the positive end of the θ continuum, closely resembling TIF curves. If we had had more informative items for the low-ability examinees, probably the scores of these examinees on the OSF would have been more similar to those of the CT and consequently the bias values would have been lower. These findings draw attention to the critical effect of the item characteristics on the results of a test, repeatedly echoed by many studies (e.g. Cella, Gershon, Lai, & Choi, 2007; Choi et al., 2010; Hol et al., 2007). The other potential source of such bias pattern is related to the

phenomenon of guessing. Previously, we discussed that due to the short length of the test, probably most of the examinees read and answered the OSF questions with more care and attention which, in turn, reduced the rate of guessing in the answers. On the other hand, due to the long length of the CT, some examinees might have lost their interest, read the questions with less care and were encouraged to guess the answers. Such examinees were logically more among low-ability examinees rather than high-ability examinees. As such, it would be then quite reasonable to expect the difference between the OSF scores and the CT scores to decrease in magnitude as θ estimates increased and approached the positive end of the continuum. It remains to note that the findings of this part must be considered in the context of the sample size limitation. Remember that for this part, we divided the sample (n = 253) into three levels. The sample size of different ability levels was unequal; particularly, the examinees of low ability level (n=9) were considerably fewer than those of the intermediate level (n=128) or high ability level (n=116). It is possible that the results of the negative end of the continuum might have differed if the tests had been administered to more low ability examinees. Thus, replication of this study with larger sample size at all ranges of ability would help to have a better understanding of the results.

As a measure of precision, the SE is an additional indicator of whether reducing the number of items substantially lowers or maintains the precision with which an OSF is estimating the ability of the test takers. As was hypothesized, the results demonstrated that OSF estimates achieved greater SE in measuring the ability of the sample than did the CT estimates. A slight loss of measurement precision in the OSF estimates is quite expected given that the length of the conventional test was substantially reduced in the OSF. The same finding was also observed in more or less all the other studies examining the psychometric quality of a short form test in comparison with a parent test (e.g. McMahon & Harvey, 2007; Wang, 2009). The results of the present study are still more encouraging than such studies as the study employed shorter forms (one-fifth of the CT) than the ones used in such studies (e.g. one-third of the CT).

The partial loss of the measurement precision in reduced test forms like OSF is inevitable particularly where the number of reduced items is substantial. In this regard, a more critical question would be to what extent the precision in scoring is required and preferable. The OSF in this study yielded precision comparable to the CATs of some studies such as Choi (2010), Hart, Mioduski, and Stratford (2005),

Thissen and Mislevy (2000), Walter et al. (2007), and Wang (2009). Occasionally, the stopping rule for a CAT program is prespecified by setting a precision standard (i.e. a standard error) in the CAT algorithm. The CAT is terminated whenever that level of standard error is met (Ware et al., 2003). The logic is that the test should not exceed that level of standard error if the test aims to be precise enough. In all the above-mentioned studies, the stopping rules of the CAT algorithms required standard errors to be less than 0.3 or 0.4 which were judged, and also showed, to be satisfactory levels of SE and led to precise ability estimates. The conclusion is that although OSF generally did not achieve the same level of measurement precision as the conventional test, it still achieved a desired and satisfied level of measurement precision and its results are comparable to those seen in a CAT system. This shows that in practice, we were able to reach adequate level of measurement precision with only 12 items, i.e. an 80% savings in the administered items and consequently in the administered time. Overall, these findings suggest that the OSF can stand as a reasonable alternative for the longer conventional test as it retained a remarkable level of precision while drastically reducing the length of the test and increasing its efficiency. Additionally, the findings suggest that the OSF can also be an efficient alternative to CAT, especially when the conditions dictate that a very short test be used but implementing a CAT program is not feasible. Of particular note is that these results were attained with a small sample size which certainly influenced the amounts of standard errors of estimations. Edelen and Reeve (2007, p. 8) maintain that "IRT scores will have smaller standard errors as sample size increases". Accordingly, it is expected that the amount of standard error of the OSF scores might be lower if the sample size were larger. Moreover, we had an item bank in which few items with high information for low-ability students were available. It would be possible that a relatively more comprehensive item bank optimizes the performance of the OSF.

The final measure of the precision of the OSF ability scores concerned the examinees' score distribution, namely ceiling and floor effects. One possible problem reported occasionally of the short form precision over the full test is related to the problem of ceiling/floor effects. For example, Choi (2010), Choi et al. (2010), Haley et al. (2004), Kosinski et al. (2003), and Ware and Sherbourne (1992) reported the presence of ceiling/floor effect in their short form tests that ranged from slight to substantial effects. This problem may come from the use of items that do not closely match the ability of the individuals at the extremes and consequently failure to capture all the levels of ability adequately (Choi, 2010). An

ideal short form therefore has enough questions to cover the entire θ range rather than limited regions of θ. In order to become as inclusive as possible, in the present study, we tried to select appropriate items at different relevant ability levels (the lower, middle, and higher ends). The results of ceiling and floor effects revealed that we were more or less successful in our goal as there were only two individuals in the ceilings but no floor effects in the OSF (despite the lack of sufficient items at the lower extreme). This would lead us to conclude that the range of *b* properly matched the range of θ in our short forms and they provided an optimal range of coverage (i.e. adequate breadth of measurement).

Comparability of θ estimates: How well do the OSF scores represent the CT scores?

Another crucial question was how well the 12-item short form scores of the OSF can stand in for the conventional test; that is how well they are comparable with and similar to the CT scores. OSF accounted for nearly a high percentage of the variance in the CT (74%), with items 5 times fewer than the CT items. This indicates that, although the short forms included only 20% of the items of the conventional test, they predicted, on average, almost 74% of the variance of the CT, suggesting that OSF could be a good predictor of the conventional test and that OSF can come quite close to the CT in terms of estimating ability. This finding mirrors the results of the previous studies on the correlation between an IRT-based short form and the full test where high correlations between the scores of the two tests were found (e.g. von Baeyer et al., 2011; Haley, Coster et al., 2004; Wang, 2009; Ware et al., 2003). This finding indicates that OSF strategy worked well with a vocabulary test since it was possible to reduce the test length as much as 80% without substantially distorting the parametric structure of the estimated scores.

A review of the short forms having nonsignificant correlations revealed that there was an item in these three short forms with item discrimination considerably lower than those of other short forms. This leads us once again to the importance of item characteristics and their influence on the results. Even one problematic item, i.e. not having high quality item parameters, may compromise the efficiency of the short forms (especially when the number of items is low), which in turn, may threaten the accuracy of the estimated scores.

Quality of classification decisions: How accurate are the results of OSF classification decisions?

Finally, an important issue regarding the utility of the OSF involves the accuracy of the classification decisions in terms of the agreement in classifications with those obtained from the CT. We observed that the OSF resulted in a classification accuracy of 91 % for the two-level classification and 87 % for the three-level classification, with the examinees having needed to answer only one-fifth of the CT items. The interesting part of the results was that for both types of classifications, the accuracy of classifications (percentage of classifications predicted correctly) increased with an increase in θ values. It was found that there were higher proportions of correctly identified master or proficient than correctly identified nonmaster or below basic level, for the two and three level classifications, respectively. This is due to the tendency that was observed for the OSF to be somewhat easier for the test takers. Recall that a substantial number of the examinees got slightly higher scores on the OSF, in comparison with the CT scores. This resulted in the identification of more examinees as master and proficient in the OSF; that is to say, in comparison with the CT, the OSF had a tendency to overidentify the level of the examinees which led to a slight decrease in the rate of correctly identifying Nonmasters and those in the Below Basic level. Nonetheless, this conclusion is true as long as the results of the CT classifications against which the OSF results were compared are correct. We earlier concluded that the negative effect of the long length of the CT might have compromised the results obtained on the CT. This is reaffirmed by Petway (2010) who notes that a measure that is too long may lead to unwanted forms of bias and threaten the clarity of the results as the examinees may experience fatigue or lose interest while responding a measure. As such, it would be possible to conclude that actually the CT had a tendency to underidentify the level of the examinees.

Overall, OSF attained classification results substantially comparable to those seen for the CT. The high degree of correspondence between the two tests in making decisions in classifying the individuals above or below one and two cutting scores is promising and encouraging. This suggests that the OSF methodology was successful in reducing the number of questions that need to be answered without seriously compromising classification accuracy; and, once more, it was indicated that the items deleted in the OSF were not informative enough and probably did little in discriminating the individuals above or below a cutting score. This high consistency of classification was a result of the advantages obtained by using the IRT model. Looking at studies that adopted more or less the same methodology, we found that Kosinski et al. (2003), for example, also found that the window

provided by IRT model contributed markedly in constructing an efficient, reliable, and valid short form for use in screening and monitoring patient outcomes.

Conclusions and Further Directions

In this study we tried to move a step forward toward developing a practical and precise shortened form of a lengthy conventional test for evaluating the vocabulary knowledge of English language learners. The findings suggest that we successfully accomplished our goal to develop an OSF that demonstrates acceptable psychometric characteristics. Accordingly, we concluded that the OSF has the potential to stand as a reasonable alternative for a longer test, especially where brevity of the test and time-saving are high priorities. The research also supported the use of the OSF as a good alternative for a CAT project, especially when conducting all those demanding steps for setting up a CAT project is not feasible, provided that items of the OSF are carefully selected on the basis of the IRT parameters. The value added by IRT was in the process taken to reach conclusions. Detailed item level information obtained from the IRT calibration allowed us to select the most appropriate items for different ability levels. Doing so, we accomplished to avoid irrelevant items for inclusion in the OSF which resulted in reducing drastically the length of the test. This, in turn, resulted in not only lightening the burden for the students, but also lessening the negative effects of a lengthy test (e.g. losing interest, guessing, cheating, and missing answers).

Concerning the partial loss of measurement precision in the OSF, it bears to mind that by and large, short tests are a tradeoff between size and accuracy. Our purpose in this study was to design short forms that conserved content coverage while sacrificing as little measurement precision as possible. However, according to Choi (2010), it is inevitable to preserve the precision in the creation of a short form. In the same vein, Petway (2010, p. 10) asserted that "any reduction will motivate a loss of some kind of information". Evidently, there is always a struggle between size and accuracy in the development of a test. The decision largely depends on which aspect, size or accuracy, of a test is the issue for a test administrator, i.e. how much information s/he is willing to obtain, and how much concern s/he has for the test size and time saving. If testing time is not strictly limited, then one may go with a CT. That is, we somewhat agree with Gosling, Rentfrow, and Swann (2003) and Rammstedt and John (2007) that static short forms are not recommended to totally substitute the regular assessments when time

and space are not in short supply or when a high degree of score precision is required.

The potential benefits of the OSFs would be admirable particularly for research contexts where routinely multiple tests as different parts of a test battery are implemented. The tests in such settings should be short enough to prevent the participants from breaking off. The availability of a short and efficient, yet psychometrically strong test is then required. Here, an OSF could be suitable to reduce the test length and decrease the threat of missing data or item selection bias without seriously compromising the accuracy of the test.

Last of all, although the findings of this study support the potential benefits of the OSFs for abbreviating the test administration process, one must weigh the pros and cons of both tests, OSF and CT, to choose the one most likely to meet the intended need. More importantly, much work still remains to be done to further our understanding of the properties of the OSF particularly in the field of language testing. Of course, the conclusions of the present study have to be interpreted in light of the specific features of the item bank, instruments, and test assembly method and cannot be generalized to other item banks, other OSF creation algorithms, or the measurement of other constructs. Likewise, the results of the classification consistency observed here should be interpreted within the specific item bank and the cut score set up here. The results of the classification errors may be different if other cut scores are set.

Notes on contributors

Faeze Safari *is currently a Ph.D. student of TEFL in the Department of Foreign Languages and Linguistics at Shiraz University, Iran. Her main interests are language assessment and second language acquisition.*

Alireza Ahmadi *is an assistant professor of TEFL in the Department of Foreign Languages and Linguistics at Shiraz University, Iran. His main interests include language assessment and second language acquisition.*

References

Ary, D., Jacobs, L. C., Sorensen, C. & Razavieh, A. (2010). *Introduction to research in education* (8th ed.). Belmont, CA: Wadsworth.

Bakeman, R., Quera, V., McArthur, D., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods, 2*, 357-370.

Cella, D., Gershon, R., Lai, J. S., & Choi, S. W. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research, 16*, 133-141.

Chang, H. H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement, 25*(4), 333-341.

Chang, H. H., & Ying, Z. L. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222.

Choi, B. (2010). *Developing precise disability measures for back pain* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3436378)

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms. *Quality of Life Research, 19*, 125-136.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cook, K. F., Choi, S. W., Crane, P. K., Deyo, R. A., Johnson, K. L., & Amtmann, D. (2008). Letting the CAT out of the bag: Comparing computer adaptive tests and an 11-item short form of the Roland-Morris Disability Questionnaire. *Spine, 33*(12), 1378-1383.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement, 20*(4), 405-416.

Doll, E. A. (1917). A brief Binet-Simon scale. *Psychological Clinic*, *11*, 197-211.

Drake, J. R. (2011). Differentiation of Self Inventory - short form: Creation and initial evidence of construct validity (Doctoral dissertation). Retrieved from https://mospace.umsystem.edu/xmlui/handle/10355/11137?show=full

Dunkel, P. A. (1997). Computer-adaptive testing of listening comprehension: A blueprint for CAT development. *The Language Teacher, 21,* 1-8.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5-18.

Giouroglou, H., & Economides, A. (2004). State-of-the-Art and adaptive open-closed items in adaptive foreign language assessment. *Proceedings of 4th Hellenic Conference with International Participation: Informational and Communication Technologies in Education* (pp. 747-756). Athens, Greece: New Technologies.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528.

Haley, S. M., Andres, P. L., Coster, W. J., Kosinski, M., Ni, P., & Jette, A. M. (2004). Short-form activity measure for post-acute care. *Arch Phys Med Rehabil, 85*(4), 649-60.

Haley, S. M., Coster, W. J., Andres, P. L., Kosinski, M., & Ni, P. (2004). Score comparability of short-forms and computerized adaptive testing: Simulation study with the activity measure for post-acute care. *Arch Phys Med Rehabil, 85,* 661-666.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hart, D. L., Mioduski, J. E., & Stratford, P. W. (2005). Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments. *Journal of Clinical Epidemiology, 58*, 629-38.

Hol, A. M., Vorst, H. C. M., & Mellenbergh, G. J. (2007). Computerized adaptive testing for polytomous motivation items: Administration mode effects and a comparison with short forms. *Applied Psychological Measurement, 31*(5), 412-429.

Jette, A. M. (2003). Assessing disability in studies on physical activity. *American Journal of Preventive Medicine, 25*(3 Suppl. 2), 122-128.

Johnson, M. A. (2006). *An investigation of stratification exposure control procedures in cats using the generalized partial credit model* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3266891)

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3315089)

Khan, A. (2010). *Use of non-parametric item response theory to develop a shortened version of the Positive and Negative Syndrome Scale (PANSS) for*

patients with schizophrenia. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3438465)

Kosinski, M., Bayliss, M. S., Bjorner, J. B., Ware, J. E. Jr., Garber, W. H., Batenhorst, A., Cady, R., Dahlof, C. G. H., Dowson, A., & Tepper, S. (2003). A six-item short-form survey for measuring headache impact: The HIT-6. *Quality of Life Research*, *12*, 963-74.

McMahon, J. M., & Harvey, R. J. (2007). Psychometric properties of the Reidenbach–Robin Multidimensional Ethics Scale. *Journal of Business Ethics, 72*, 27-39.

Padaki, M., & Natarajan, V. (2009). An approach to implementing adaptive testing using item response theory both offline and online. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Petway, K. T. (2010). *Applying adaptive methods and classical scale reduction techniques to data from the big five inventory* (Master's thesis). Available from ProQuest Dissertations and Theses database. (UMI No. 1479936)

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*(1), 203-212.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*(4), 347-364.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102–111.

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55*(1), 167-194.

Thissen, D., & Mislevy, R. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed.) (pp. 101-133). Hillsdale, NJ: Lawrence Erlbaum Associates.

von Baeyer, C. L., Chambers, C. T., & Eakins, D. M. (2011). Development of a 10-item short form of the Parents' Postoperative Pain Measure: The PPPM-SF. *Journal of Pain*, *12*(3), 401-406.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for Anxiety (Anxiety-CAT). *Quality of Life Research, 16*(1), 143 -55.

Wang, J. H. (2009). *Using real-data simulations to compare computer adaptive testing and static short-form administrations of an upper extremity item bank*

(Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3386014)

Ware, J. E. Jr., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Dahlöf, C. G., Tepper, S., & Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12*, 935-952.

Ware, J. E. Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): I. conceptual framework and item selection. *Medical Care, 30*(6), 473-483.

Weiss, D. J. (2004). Computerized adaptive testing for Effective and Efficient Measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*, 70-84.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing, 27*(1), 119-140.

Zhang, Y. (2006). *Impacts of multidimensionality and content misclassification on ability estimation in computerized adaptive sequential testing (CAST)* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3221052)

Zimowski M., Muraki E., Mislevy, R., & Bock, R. D. (2003). BILOG-MG. In M. du Toit (Ed.), *IRT from SSI* (pp.24-256). Lincolnwood, IL: Scientific Software International, Inc.