LANGUAGE
TESTING

Language Testing
27(1) 119–140
© The Author(s) 2010
Reprints and permission: http://www.sagepub.co.uk/journalsPermission.nav
DOI: 10.1177/0265532209347363
http://ltj.sagepub.com
SAGE

Assessing the accuracy and consistency of language proficiency classification under competing measurement models

Bo Zhang
University of Wisconsin – Milwaukee, USA

Abstract

This article investigates how measurement models and statistical procedures can be applied to estimate the accuracy of proficiency classification in language testing. The paper starts with a concise introduction of four measurement models: the classical test theory (CTT) model, the dichotomous item response theory (IRT) model, the testlet response theory (TRT) model, and the polytomous item response theory (Poly-IRT) model. Following this, two classification procedures are presented: the Livingston and Lewis method for CTT and the Rudner method for the three IRT-based models. The utility of these models and procedures are then evaluated by examining the accuracy of classifying 5000 language test takers from a large-scale language certification examination into two proficiency categories.

The most important finding is that the testlet format (multiple questions based on one prompt), which language tests usually rely on, has a great impact on the proficiency classification. All testlets in this study show a strong testlet effect. Hence, the TRT model is recommended for proficiency classification. Using the standard IRT model would inflate the classification accuracy due to the underestimated measurement error. Meanwhile, using the Poly-IRT model would give slightly less accurate classification results. Concerning the CTT model, while its classification accuracy is comparable to that of the TRT, there exists considerable inconsistency between their classification results.

Keywords

item response theory, language proficiency, language testing, proficiency classification, testlet

Proficiency classification has played a vital role in second language testing. It is one of the major, if not the only, reasons for which many language learners actually take language tests. Most large-scale standardized tests of English as a second language, such as the Test of English as a Foreign Language (TOEFL), the International English Language Testing Service (IELTS), and the Michigan English Language Assessment Battery

Corresponding author:
Bo Zhang, Department of Educational Psychology, University of Wisconsin – Milwaukee, P. O. Box 413, Milwaukee, WI 53201–0413, USA.
E-mail: boz@uwm.edu

(MELAB), serve for classifying examinees to some degree. In using the scores from these tests, universities and colleges usually set up a minimum score to classify their applicants as those who meet the language requirement and those who don't. By nature, classification decisions are of high risk. Errors may result in individuals being deprived of well-deserved educational or career development opportunities. Unfortunately, like in any other educational test, it is almost impossible to avoid measurement error in estimating the proficiency levels in language tests. Consequently, classification errors are also inevitable; hence evaluating the accuracy of test scores used to represent proficiency categories is of great importance.

Language proficiency refers to a person's general communicative competence in the target language environment (Canale and Swain, 1980). The exact nature of language proficiency or language ability has undergone some dramatic changes over the past few decades. In general, the assumption that language ability is a 'unitary competence' (Oller, 1979) has gradually been replaced by the belief that language competence is more complex and consists of multiple inter-correlated abilities and strategies (Bachman, 1991). One representative example of this multi-component structure is the three-tier hierarchical model proposed by Bachman and Palmer (1996). According to this model, top tier consists of language knowledge and strategic competence. At the second tier, the knowledge component can be further divided into organizational knowledge and pragmatic knowledge. Meanwhile, strategic competence is composed of strategies used in goal setting, assessment, and planning. Finally, at the bottom tier, organizational knowledge can be expressed as either grammatical knowledge or textual knowledge, while pragmatic knowledge encompasses functional or sociolinguistic knowledge. According to this model, a proficient language speaker should not only demonstrate the structural knowledge of a target language but should also have the necessary strategies to implement that knowledge effectively in actual use.

On the other hand, most language teachers are more familiar with the traditional definition whereby language proficiency comprises linguistic skills in the four core curricular areas: listening, speaking, reading, and writing. While this conceptualization may facilitate everyday language instruction, one possible drawback is the over-generalization of these basic skills (Bachman and Palmer, 1996). That is, tasks vastly different in nature may be classified under the same category. For example, listening to someone talk in person and getting ready to respond is very different from listening to a news announcement on the radio, but both would be labeled as involving listening comprehension proficiency under the above traditional definition.

As far as proficiency classification is concerned, language instructors and testers have multiple options. With regard to measurement, they may use models based on either classical test theory (CTT) or item response theory (IRT). IRT itself provides a number of models for analyzing any single test. Meanwhile, these models may be applied at different levels of a test. For example, if multiple components of a test can be represented by a meaningful unified score, proficiency classification may be conducted at the test level. On the other hand, there may be interest in classifying examinees according to separate curricular areas so that more diagnostic information may be obtained.

Faced with these options, practitioners should be informed of the results and consequences of different procedures in order to choose the one that best satisfies their needs. Most importantly, they should be guided to find the procedure that minimizes proficiency classification errors. In reality, as relevant research on language proficiency classification is extremely scarce, this type of information is very limited. While a number of proficiency classification methods have been proposed and evaluated for general use (e.g. Hanson and Brennan, 1990; Livingston and Lewis, 1995; Rudner, 2001; Wainer et al., 2005), none of them has been systematically studied for use in language tests. As a result, it is unclear how these procedures may be applied to various language testing conditions.

The main purpose of this study is to gather empirical evidences to help practitioners undertake appropriate proficiency classification in language testing. Two objectives guide this research. The first is to evaluate classification accuracy under the aforementioned four measurement models. Clearly, the fewer classification errors a model makes, the more valuable it is. The second objective is to study the consistency of classification results when different measurement models are applied. In particular, results from the testlet response theory (TRT) TRT model are compared to those from other models.

This paper begins with a concise introduction to the theoretical framework of the four measurement models. Following this, two classification procedures are presented: the Livingston and Lewis (1995) method for CTT and the Rudner (2005) method for the three IRT-based models. Then, the efficacy of these classification procedures is evaluated by using data from a large-scale certification test. Finally, the implications of the findings from this research for the classification of overall language proficiency are discussed.

Competing measurement models

Classical Test Theory (CTT) Model

Classical test theory, also known as the true score test theory, assumes that any obtained test score is a sum of two elements: the true ability that has motivated the measurement, and the measurement error that is almost ubiquitous in educational testing. The CTT model is simply expressed as (Allen and Yen, 1979):

$$X = T + E \tag{1}$$

where X is the observed score, T is the true score, and E is the error score. In any measurement, only the observed score is known. To estimate the true score, some strong assumptions have to be made. Under CTT, it is assumed that measurement error is random and the true score is the expected value of the observed score. In other words, the true score for an examinee is the average of the observed scores from an infinite number of measurements of this examinee. Under this assumption, the exact value of any true score can never be estimated and, thus, the true score under CTT remains a theoretical construct.

Test reliability under CTT is defined as:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}, \tag{2}$$

where X, T, and E have been defined in Equation 1, ρ_{XX} is the reliability coefficient, σ_r^2 is the true score variance, and σ_X^2 is the observed score variance. To estimate the magnitude of measurement error, the standard error of measurement (SEM) is expressed as:

$$SEM = \sigma_X\sqrt{(1 - \rho_{XX'})}, \tag{3}$$

where σ_X is the standard deviation of the observed score. Once the *SEM* is known, with the assumption that X is a random variable with a mean of T and a standard deviation of *SEM*, a confidence band may be built to estimate the true score.

The advantages of conducting proficiency classification under CTT are obvious. As shown in Equation 1, the measurement model is relatively simple and a variety of methods have been developed for estimating test reliability, such as by coefficient alpha (Cronbach, 1951) and split-half reliability (Spearman, 1910). These statistics have also been incorporated into standard statistics software packages, such as SAS and SPSS. The disadvantages of using CTT, on the other hand, are also clear and somehow insurmountable. As true ability is known through a confidence interval only, it has to be approximated in the proficiency classification, which is likely to increase classification error. Moreover, the SEM as computed in Equation 3 relies on information at the test level (i.e. the standard deviation of the observed score distribution and test reliability), hence its value will be constant across all examinees. As the SEM generally varies across the range of individual proficiencies (e.g. Peterson et al., 1989), this classical SEM index provides only an estimate of the average measurement error for all examinees. Kolen, Hanson, and Brennan (1992) thus suggested that the SEM conditional on the proficiency level, or the conditional SEM, should be estimated. To obtain a point estimate of the true proficiency level along with the person-specific *SEM*, one has to resort to models based on item response theory.

Dichotomous Item Response Theory (IRT) Model

Item response theory (Lord, 1980) has gradually become the mainstream theory in educational measurement. It is currently applied in most large-scale standardized achievement tests (e.g. SAT, ACT, GRE, LSAT, MCAT, and Melab) as well as most state accountability tests. IRT models reflect the interaction between test items and test takers by means of a probabilistic relationship. The most commonly used IRT models are the unidimensional logistic models for scoring dichotomous items. The three-parameter logistic model, or the 3PL, is expressed as (Birnbaum, 1968):

$$P(Y_{ij} = 1|\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i - \lambda_{id(j)})}}, \tag{4}$$

where p is the conditional probability that the response Y_{ij} from person j to item i is correct, θ is the underlying proficiency or ability level, c is the item guessing parameter, *a* is the item discrimination parameter, *b* is the item difficulty parameter, and *D* is a scaling factor. Note that the term 'ability' is used interchangeably with the term 'proficiency' in this writing as they both refer to the underlying linguistic competence which a test is designed to assess.

For items with no chance of being guessed correctly (e.g. short-answer items with correct/incorrect scoring), the *c* parameter would drop from Equation 4 and the model would reduce to the two-parameter model. If the discrimination parameter can be further assumed to be constant across all items, the one-parameter IRT model with the item difficulty parameter only may be applied. When the discrimination parameter is fixed at 1 for all items, the one-parameter model reduces to the Rasch model, which is probably the most widely used IRT model in language testing to date (e.g. Adams et al., 1987; Lynch et al., 1988; McNamara, 1990).

Differing from the CTT, in which true ability can only be known through a confidence interval, using IRT will provide a point estimate of each examinee's true proficiency. This estimation is quite independent of the particular choices of specific test items from the potential population of all items: the so-called test-free measurement property (Hambleton and Swaminathan, 1985). In addition, as the person-specific standard error can be estimated, the likelihood that a positive or negative error has been committed in the proficiency classification can be evaluated more accurately under IRT.

Testlet Response Theory (TRT)

An important assumption under the IRT model is local independence (Hambleton and Swaminathan, 1985). This assumption states that the relationship among items in any test is established through nothing but the measured ability. For any individual test taker, a response to any item should not be affected by responses to any other items. In other words, responses should be independent. This assumption can also be expressed as follows: no ability dimension other than the targeted one should have affected item responses.

A common condition that may indicate the local independence assumption has been violated is the application of testlets (e.g. Rosenbaum, 1988; Yen, 1993). A testlet is defined as a group of items based on the same stimulus (Wainer and Kiely, 1987). Testlets are commonly employed in language assessments. A classic example of a testlet is a reading passage followed by a number of multiple-choice questions. Responses to all items in such a testlet not only depend on reading competence but also on the understanding of specific contextual or cultural background information embedded in the common stimulus. For students with insufficient background knowledge, it is likely that responses to all items in the testlet would be affected. Using IRT terminology, these items are locally dependent.

When the local independence assumption is untenable, using the standard IRT model would not provide the appropriate interpretation of test results because the model would no longer fit the responses. Specifically, the item discrimination parameter would be overestimated (Yen, 1993). As the discriminating power of test items represents the amount of information an item contributes to ability estimation, the overall test information is also likely to be overestimated (Sireci et al., 1991; Thissen et al., 1989). Hence, the major harm that local dependence (LD) does to IRT modeling is the inflation of measurement precision.

One direct way to handle the LD effect is by adding a testlet effect term to the IRT model. A testlet response (TRT) model (Bradlow et al., 1999) is formulated as:

$$P(Y_{ij} = 1|\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i - \lambda_{id(j)})}} \tag{5}$$

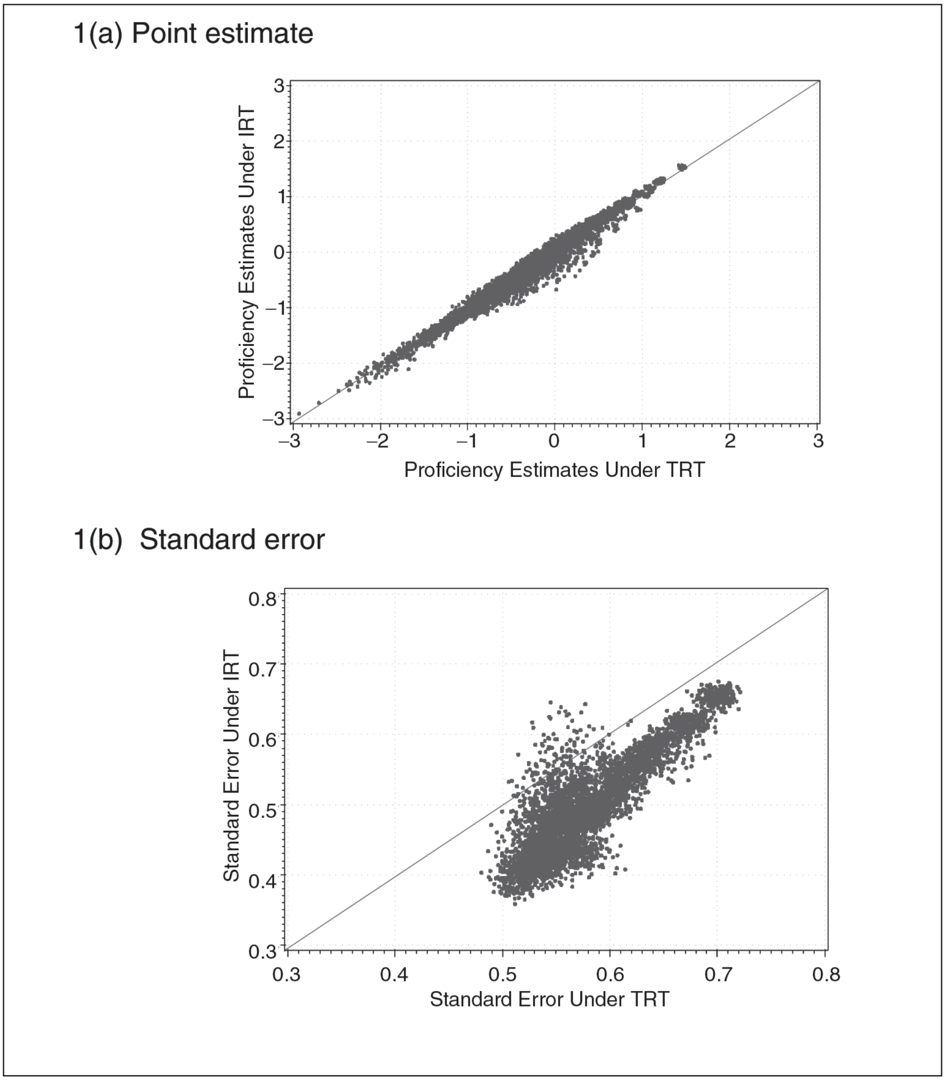
Compared to the standard IRT model as described in Equation 4, the only new term here is $\lambda_{id(j)}$, which is the testlet effect for person *j* in answering item *i* nested within testlet *d*. The term $\lambda_{id(j)}$ is assumed to be centered around 0. Its variance indicates the severity of any local dependence.

In Figure 1, the test information inflation due to LD items is illustrated by the reading test in the Examination for the Certificate of Proficiency in English (ECPE). More details of this examination can be found in the Method section of this article. In this reading test, examinees read four paragraphs, each followed by five multiple-choice items. As questions about the same paragraph share the same stimulus, they are locally dependent. The IRT estimates in the figure represent estimates derived by applying the standard 3PL model. In this case, any possible LD effect has been completely ignored. In Figure 1a, not much difference was observed between the point estimates of true proficiency by these two models. Most dots in the figure are close to the 45-degree reference line, indicating that estimates from these two models are about equal. However, in Figure 1b, the standard error of the ability estimates from the TRT model is larger than that from the IRT model for most examinees. What this means is that if the IRT model were selected for the proficiency estimation of this reading test, test users would be overconfident about their measurement precision as the IRT model would show less measurement error than is, in fact, the case. Note that in Figure 1b, a very small portion of examinees actually shows larger standard error under IRT than under TRT. This inconsistency may be attributed to random measurement error.

Polytomous Item Response Theory (Poly-IRT) Model

Another possible way to handle the LD effect is by using the polytomous IRT model (Thissen et al., 1989). This method first collapses the responses from locally dependent items into a polytomous item, thus eliminating any possible LD effect. Next, a polytomous item response theory model will be applied to obtain a proficiency estimation. A popular model for polytomous items is the graded response model, proposed by Samejima (1969). This model takes a two-step approach in modeling how an examinee responds to a polytomously scored item. The first step is to compute the conditional probability that examinee *j* will score in the response category *k and higher* in item *i* by the following function:

$$P_{ijk} * (\theta) = \frac{1}{1 + e^{-a_i(\theta_j - b_{ik})}}, \tag{6}$$



1(a) Point estimate

1(b) Standard error

Figure 1. Comparing the proficiency estimates from the TRT and IRT models

Proficiency estimates from the ECPE reading test were first obtained by using the IRT model, then by the TRT model. In this figure, the differences from using these two models were depicted by the point estimate in 1(a) and the standard error in 1(b).

where P_{ijk}* is the conditional probability, b_{ik} is the step difficulty, and all other terms share the same interpretation as in Equation 4. Next, the conditional probability for the score category k is the difference between the conditional probability of two adjacent categories:

$$P_{ijk}(\theta) = P_{ijk} * (\theta) - P_{ij(k+1)} * (\theta). \tag{7}$$

While the collapsing of locally dependent items will eliminate the LD effect, one potential drawback of this practice is the loss of test information (Yen, 1993). This loss tends to be particularly severe for testlets with a large proportion of locally dependent items. Take, for example, the cloze test in the present study which has 20 items, all based on the reading of one passage. Proficiency estimation using the polytomous IRT model for this test would rely on responses to only one item with 21 categories (0–20). In other words, using the polytomous model may not represent accurately how examinees have responded to the original 20 dichotomous items.

Proficiency classification

Proficiency classification accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error (Hambleton and Novick, 1973). In educational testing, this accuracy must be estimated because errorless test scores never exist. Any misclassification of an examinee would indicate a classification error. A false positive error occurs when an examinee is classified at a higher proficiency category than the true one, whereas a false negative error results when an examinee is put into a category lower than their true ability. In practice, which type of error is of more concern is a matter of judgment.

One straightforward method to measure classification accuracy is through comparing the classification results based on scores from two equivalent forms of the same test. If examinees are classified consistently into the same categories by both forms, classification accuracy is high. The challenge of this method lies in the difficulty of justifying testing the same examinees twice using the same test. Accordingly, classification accuracy has to be evaluated based on one single test administration. A number of such procedures have been developed, some based on CTT (Hanson and Brennan, 1990; Huynh, 1976; Lee et al., 2004; Livingston and Lewis, 1995; Subkoviak, 1976) and others on IRT (Rudner, 2005; Wainer et al., 2005).

To evaluate classification accuracy under CTT, a true score distribution needs to be approximated. The present study employed the procedure developed by Livingston and Lewis (1995) (hereafter referred to as LL) for the CTT classification. This method assumes that the proportional true score follows a four-parameter beta distribution. Based on the first four moments of the observed score distribution, the exact form of the true score distribution may be estimated by a method proposed by Lord (1965). Once a true score distribution is defined, an assumed score distribution from an alternate form can be estimated. The LL procedure compares the observed score distribution to the reconstructed alternate score distribution in order to assess the classification accuracy (Brennan, 2004). For the exact steps and technical details of the LL method, refer to Livingston and Lewis (1995). With regard to the effectiveness of the LL procedure, Wan, Brennan, and Lee (2007) conducted a simulation study and concluded that the LL procedure yielded relatively accurate decision results, compared to four other classification methods under CTT.

Under the IRT framework, the point estimate of ability may be treated as the true score on the latent trait. Thus, the approximation of the true score distribution in CTT is

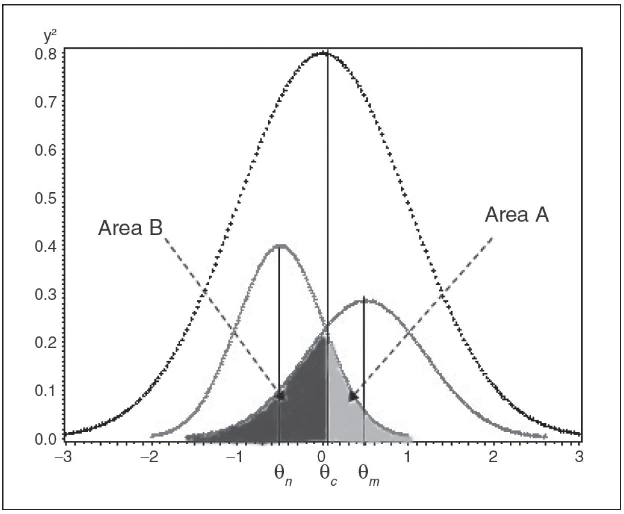


Figure 2. Illustration of the false positive and negative errors in proficiency classification

In this figure, θ_c is the cut-off score, θ_n and θ_m are the true proficiency levels for two examinees. Area A contains the estimates of θ_n that are larger than the cut score, thus the positive errors. Likewise, Area B represents the negative errors.

unnecessary. The major challenge shifts to how to account for the measurement error associated with the point estimate of proficiency levels. Rudner (2001, 2005) introduced a method for evaluating the decision accuracy through the computation of the expected likelihood of classifications. In the following, without loss of generality, this method will be described using a pass/fail classification scheme.

Suppose that the passing score is θ_c. The true ability is θ_n for Examinee A and θ_m for Examinee B. Their positions on the ability scale are depicted in Figure 2. Due to the error associated with the ability estimation, a conditional distribution accompanies each true theta. As θ_n is smaller than θ_c, Examinee A should be classified as a non-master in all estimations. Likewise, Examinee B should be a master. However, there is a clear chance that Examinee A would be classified as a master. That chance can be represented by the size of Area A in the figure, where the theta estimates are larger than the cut score θ_c. In classification terminology, this chance is the likelihood that a false positive error would be committed, whereby a true non-master is identified as a master.

The size of Area A can be computed as the area to the right of the following z score:

$$z = \frac{\theta_c - \theta_n}{se(\theta_n)}, \tag{8}$$

where $se(\theta_n)$ is the standard error of the θ_n estimates. The expected frequency of false positive errors for all examinees equals the sum of the above likelihood over all non-masters, or

$$L(m,n) = \sum_{n=1}^{N} [p(\hat{\theta}_n > \theta_c | \theta_n) f(\theta_n)], \tag{9}$$

where $L(m,n)$ refers to the frequency that non-masters are classified as masters, N is the number of non-masters, $\hat{\theta}_n$ is the θ_n estimate, and $f(\theta_n)$ is the population density of θ_n. Likewise, the frequency of false negative errors (masters classified as non-masters) can be calculated by

$$L(n,m) = \sum_{n=1}^{M} [p(\widehat{\theta}_m < \theta_c | \theta_m) f(\theta_m)], \tag{10}$$

where M is the number of masters. The expected frequencies for the correct classifications for both masters and non-masters could be computed in the same manner as in the last two equations. In order to evaluate the classification accuracy, these expected frequencies can then be compared to the observed.

Using the Z score to compute the probability in Equation 8 relies on the assumption of normality of the conditional distribution of the theta estimates. Guo (2006) introduced a method based on the likelihood function of the ability estimates which is thus free from the assumption of normality. For testing conditions examined in that study, results with or without the assumption of normality were similar. Wainer et al. (2005) studied the proficiency classification under the Bayesian framework. As the Markov chain Monte Carlo (MCMC) procedure was employed for the proficiency estimation, the exact values of the conditional distribution of the ability estimates are available. Accuracy of proficiency classification could be assessed by simply counting the number of times that false positive or false negative errors have been committed.

Method

Instruments and participants

The Examination for the Certificate of Proficiency in English (ECPE) is an English language proficiency test for adult non-native speakers of English at the advanced level (English Language Institute, 2006). Learners take this test to be certified as having the necessary English skills for education, employment, or professional business purposes. The test assesses English language proficiency in the following areas: speaking, writing, listening, cloze, grammar, vocabulary, and reading. In reporting, grammar, cloze, vocabulary, and reading are scored together as one section labeled as the GCVR test. Candidates must pass all four sections in order to be awarded the certificate.

This study investigated proficiency classification for two sections in the ECPE: listening and GCVR. Both sections have a large number of items, allowing investigation of the efficacy of different measurement models in proficiency classification. The 50-item listening test consists of two parts. The first 35 items are independent items, each based on an independent prompt. The last 15 items are based on three long dialogues or paragraphs, each followed by five questions. These items are locally dependent and

Table 1. Measurement models for various tests

Subtests	Measurement models			
	CTT	IRT	Poly-IRT	TRT
Listening	x	x	x	x
GCVR	x	x	x	x
Grammar	x	x		
Vocabulary	x	x		
Reading	x	x	x	x

susceptible to the testlet effect. In the cloze test, all 20 items share one stimulus, thus a strong testlet effect may be present. The reading test asks examinees to read four paragraphs, each followed by five questions. Again, it is highly possible that these items will show local dependence. The grammar and vocabulary tests each use 30 independent items. Subjects were 5000 examinees, randomly selected from a one-year administration of the ECPE.

Measurement models and model estimation

Table 1 lists the measurement models applied to each test. The grammar and vocabulary tests use no testlet, thus the results from TRT would be equivalent to those from IRT. The listening, GCVR, and reading tests all showed strong testlet effects. The appropriate model for them is thus the TRT model. The IRT model was applied to investigate possible damage to the proficiency classification should local dependence be ignored.

As all items were multiple-choice items, the three-parameter logistic model was applied for both IRT and TRT. The proficiency estimation under IRT was obtained by using the MULTILOG computer program (Thissen, 1991). This program implements the marginal maximum likelihood method to estimate the ability trait. To increase the estimation accuracy, prior distributions were imposed on item parameters as follows: normal (1.1, 0.6) for the a's, standard normal for the b's, and normal (−1.1, 0.5) for the logit form of the c's. As the sample size of this study was large (i.e. 5000), the impact of these priors on a's and b's was probably quite limited (Harwell and Janosky, 1991). The main purpose of using these priors was to constrain the c parameter to reasonable values. The proficiency estimation for the Poly-IRT model was also conducted by using the MULTILOG program.

For the TRT model, parameter estimation was based on the Markov chain Monte Carlo (MCMC) procedure, as operationalized in the Scoright program (Wang et al., 2004). This program adopts the full Bayesian structure for estimating testlet model parameters. For details on the estimation algorithm, refer to Wang, Bradlow, and Wainer (2002), or the Scoright program manual (Wang et al., 2004). One important issue in the MCMC estimation is monitoring the convergence of the posterior distribution of model parameters. Following the suggestions in the Scoright manual on how to improve and check the model convergence, a potential scale reduction factor close to 1 was set as the convergence criterion. In addition, three chains were run, each thinned with five draws to reduce the autocorrelation effect. Convergence was achieved for all tests.

Cut scores for proficiency classification

As the ECPE test serves a certificatory purpose, classification decisions based on its results are binary by nature. Accordingly, cut scores in this research were used to separate examinees into two categories: masters versus non-masters. The following steps were taken to establish cut scores that are comparable across different measurement models. First, in the year that the test data were sampled, 52% of examinees passed the listening test and 58% passed the GCVR. One can reasonably assume that these percentage groups of examinees were masters under all the investigated models. Next, for each measurement model, the estimated proficiency level that corresponded to the above percentile ranks (i.e. 48 for the listening test, 42 for the GCVR and its subtests) were set as cut scores.

Results

Before proficiency classification was conducted, how well each test item measured the relevant language proficiency was examined. This is important as items that do not measure the corresponding trait properly may invalidate the application of measurement models. For this purpose, the corrected point-biserial correlation between item responses and corresponding section total score was first computed. The term ‘corrected’ implies responses to the item under study were not included in the computation of the total score. Three items – two from the listening test and one from the vocabulary test – showed a negative correlation. The IRT analysis also indicated that these items in this sample had a negative discrimination parameter, implying that examinees with higher ability were actually less likely to answer these items correctly. Consequently, these items were excluded from the proficiency classification.

Table 2 presents Cronbach’s alpha coefficients for the reliability of the tests under study. The listening and GCVR tests showed the highest reliability, which is important as their classification results are actually reported in practice. The grammar, vocabulary, and reading tests all had reliability around 0.7. While 0.7 has been recommended as a general guideline for acceptable reliability (Nunnally and Bernstein, 1994), for high-stakes decisions such as proficiency classification, a higher reliability coefficient is more desirable. Not surprisingly, coefficient alpha is lowest for the cloze test. This test has the shortest length and, furthermore, cloze tests generally measure a variety of linguistic skills at both the syntax and discourse levels (Alderson, 1979; Bachman, 1982; Oller, 1973; Saito, 2003 specifically for the ECPE test), making them susceptible to showing low coefficient alpha. Moreover, the cloze test is more a test format than a curricular area or a language competence. For all these reasons, proficiency classification was not performed for the cloze test in this study.

Table 2. Test reliability coefficient

Subtests	No. of items	Reliability
Listening	48	0.76
Grammar	30	0.70
Cloze	20	0.58
Vocabulary	29	0.73
Reading	20	0.74
GCVR	99	0.86

Table 3. Magnitude of the testlet effect

Subtest	No. of testlets	Testlet effect
Listening	1	1.05
	2	0.41
	3	1.09
Cloze	1	1.43
Reading	1	0.58
	2	0.53
	3	0.35
	4	0.59
GCVR	1	0.27
	2	1.31
	3	1.44
	4	1.33
	5	1.83

Data on the magnitude of the testlet effect is presented in Table 3. Following the suggestion by Bradlow, Wainer, and Wang (1999) that a variance over 0.3 for the testlet term $\lambda_{id}(j)$ indicates sizable testlet effect, all tests relying on testlets demonstrated strong testlet effects, which also confirms the belief that the reading and listening passages in the ECPE test violate the local independence assumption. As expected, the cloze test demonstrated strong testlet effect since all items in this subtest shared the same prompt. Note that same testlets, when placed in different tests, exhibited different magnitude of the testlet effect. As an example, the cloze items demonstrated weaker testlet effect in the GCVR test than in the cloze test. But the testlet effect for the reading items became stronger in the GCVR test than in the reading test. While it is hard to illuminate this change of effect by studying the item responses only, analyzing the content of these testlets may shed some light on these shifts.

Table 4 gives the summary statistics for the proficiency estimate under the four measurement models. Note that the raw score and the three IRT-based estimates are not on the same scale, thus their means and standard deviations (SD) should not be compared. In addition, the IRT scales were centered around 0 during the estimation, hence, their means should all have been 0. Any non-zeros are due to the estimation error. As shown in the table, the proficiency distribution by IRT and TRT are very similar for the listening

Table 4. Summary statistics of the estimated proficiency under different models

		Mean	SD	Skewness	Cut score
Listening	IRT	0.00	0.89	–0.07	–0.016
	Testlet	0.00	0.88	–0.09	–0.017
	Poly-IRT	0.01	0.85	–0.08	–0.013
	CTT	33.90	5.79	–0.33	34
GCVR	IRT	0.00	0.94	0.08	–0.189
	Testlet	0.00	0.93	0.10	–0.194
	Poly-IRT	0.04	0.91	0.08	–0.144
	CTT	67.19	11.40	–0.16	65
Reading	IRT	0.00	0.86	–0.20	–0.173
	Testlet	0.00	0.81	–0.23	–0.144
	Poly-IRT	–0.02	0.79	–0.26	–0.152
	CTT	15.01	3.35	–0.69	15
Vocabulary	IRT	0.00	0.87	0.01	–0.254
	CTT	17.05	4.54	–0.02	14
Grammar	IRT	0.00	0.85	–0.06	–0.167
	CTT	21.70	4.10	–0.41	22

and GCVR tests. One major difference is the standard deviation of the Poly-IRT estimates is consistently smaller than that of IRT and TRT. In other words, collapsing items in testlets into polytomous items tend to shrink the proficiency difference among the examinees. Meanwhile, the skewness statistics indicate that the IRT-based estimates are distributed more symmetrically than the raw score.

Figure 3 illustrates different proficiency distributions for the listening test. The shape of these distributions is very similar and very close to the normal distribution. As the GCVR test shows the same pattern, to save space, its histograms are not presented. Instead, results for the reading test are given as more difference was observed. In Figure 4, while the proficiency distributions under IRT and TRT are still alike, the Poly-IRT and raw score distributions look quite different. Specifically, the Poly-IRT distribution is considerably less even. For example, there are about 7% of examinees at the proficiency category of 0.5 but only 2% at the 0.2 level. In contrast, the corresponding percentage is about 5% at both levels under IRT and TRT. Meanwhile, the raw score distribution is apparently more negatively skewed than the other three.

Table 5 presents the classification results for listening and GCVR tests. Cell values indicate the percentage of examinees falling into each category. The column labeled as ‘accuracy’ gives the percentage of examinees expected to be classified correctly. Using the listening test as an example it can be seen that under the testlet model, 41.5% of examinees had been correctly identified as not passing and 43.3% as passing. Thus 84.8% (41.5% plus 43.3%) of total examinees were expected to be correctly identified. Meanwhile, 7.5% of examinees could be misclassified as masters. They represented the false positives. False negatives are the 7.7% of examinees who were expected to pass but actually were classified as failing.

The overall classification accuracy was high for both tests. Results based on different measurement models were similar. For the listening test, about 85% of

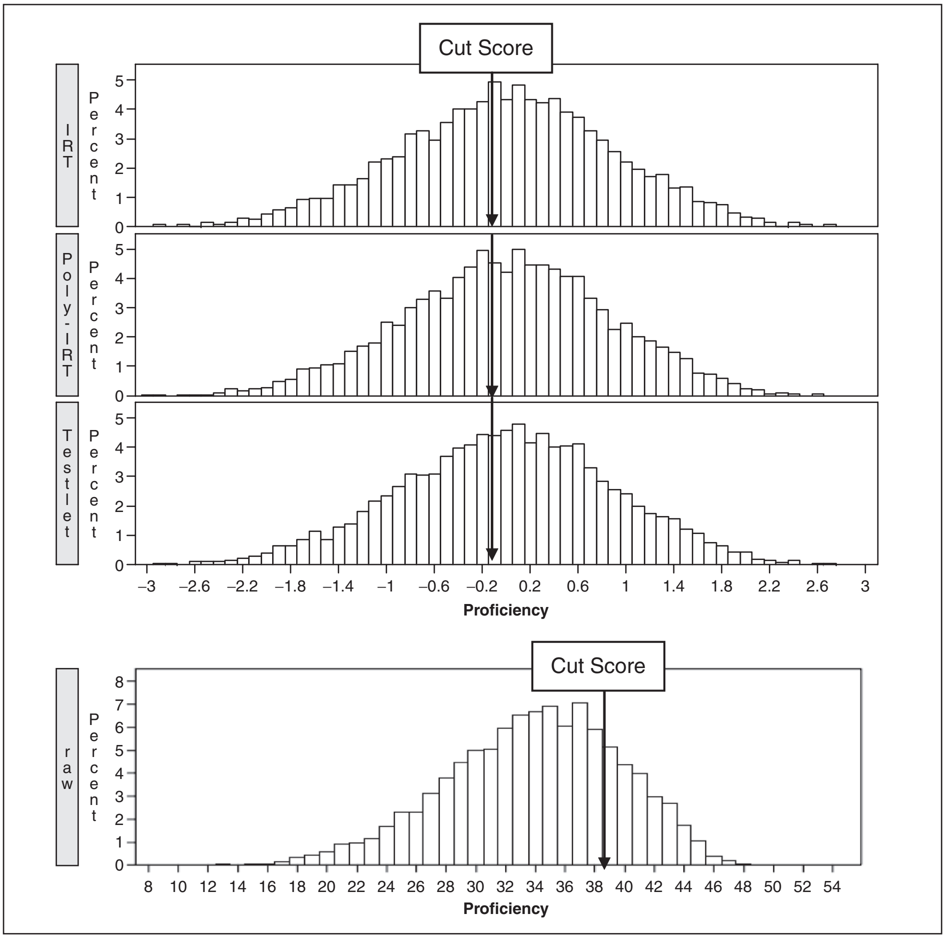


Figure 3. Illustration of the proficiency distribution under four models for the Listening Test

In this figure, the x-axis is the proficiency estimate from each model and the y-axis indicates the percentage of examinees falling into each proficiency category. Area to the right of the cut score line represents the total percentage of the masters. As the cut-score for the three IRT models are very close to each other, their lines almost overlapped.

examinees were expected to be correctly identified, which also means about 15% of examinees may be misclassified. Classification accuracy was slightly higher for the GCVR test with the percentage of correct identification around 87%. For both the listening and GCVR tests, the false positive error rate and the false negative error rate were about equal.

Table 6 gives the results for the subtests under GCVR. Classification accuracy was clearly lower for these tests and both false positive and false negative errors increased considerably. For the reading test, the classification error rates were considerably lower under IRT than under TRT. This seemingly higher accuracy should not be interpreted as

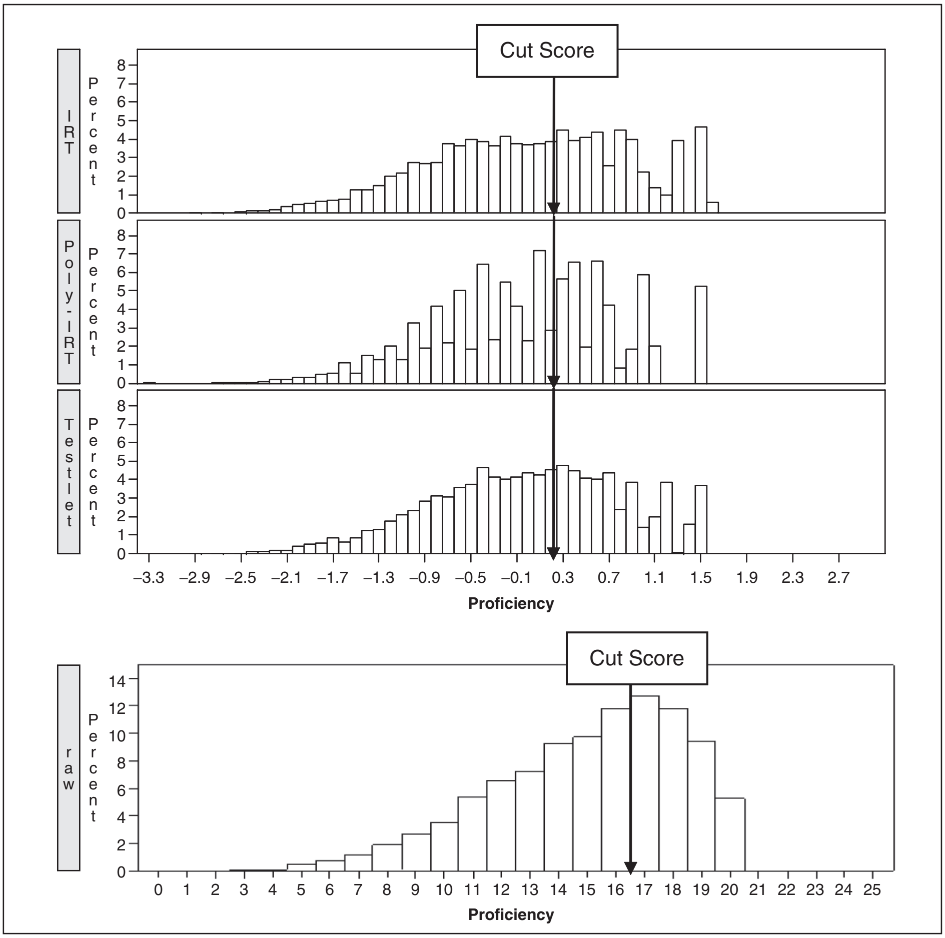


Figure 4. Illustration of the proficiency distributions under four models for the Reading Test
This figure shows how the proficiency distribution under Poly-IRT and CTT is different from that under IRT and TRT.

meaning that the IRT model would provide more accurate proficiency classification results. As shown in Table 3, the reading test had a strong testlet effect, thus the accuracy of proficiency estimation had been inflated under the IRT model. What this table actually reveals is that using the IRT model would also overestimate the accuracy of proficiency classification.

Finally, the consistency between the TRT model and the other three models in terms of proficiency classification are reported in Table 7. For all these tests, the agreement between the TRT and IRT models was above 95%. This is even true for those tests with a strong testlet effect. This high agreement was not surprising as the point estimates of the proficiency level were similar under these two models, as exemplified by the reading test in Figure 1.

Table 5. Classification accuracy: Listening and GCVR tests

Tests	Measurement models	Classified proficiency levels	Expected proficiency levels		Accuracy
			Fail	Pass	
Listening	Testlet Model	Fail	41.5	7.7	
		Pass	7.5	43.3	
					84.8
	IRT Model	Fail	41.5	7.6	
		Pass	7.5	43.5	
					85.0
	Poly-IRT Model	Fail	41.3	8.0	
		Pass	7.7	43.0	
					84.3
	CTT	Fail	39.8	8.1	
		Pass	7.7	44.3	
					84.1
GCVR	Testlet Model	Fail	37.0	5.8	
		Pass	6.0	51.1	
					88.1
	IRT Model	Fail	37.1	5.6	
		Pass	5.9	51.4	
					88.5
	Poly-IRT Model	Fail	36.8	5.9	
		Pass	6.2	51.1	
					87.9
	CTT	Fail	35.8	6.0	
		Pass	5.6	52.6	
					88.4

The Poly-IRT model provided comparable results to the TRT model for listening and GCVR tests. The agreement between these two models was considerably lower for the reading test, probably due to the fact that all testlet items were collapsed into polytomous items. For listening and GCVR tests, the CTT model tended to classify more masters under TRT as non-masters than vice versa. However, the reading test shows an opposite pattern in that as many as 8% of non-masters under IRT would be identified as masters under CTT.

Discussion

The Examination for the Certificate of Proficiency in English (ECPE) was presented above as a representative example of commonly used second language tests. Such tests measure general language competence by assessing skills in the key curricular areas such as listening, reading, grammar, vocabulary, and writing. They routinely use long passages as the prompts in order to ask a large number of multiple-choice items. The main

Table 6. Classification accuracy: Grammar, vocabulary, and reading tests

Tests	Measurement models	Classified proficiency levels	Expected proficiency levels		Accuracy
			Fail	Pass	
Grammar	IRT Model	Fail	35.0	8.7	
		Pass	8.0	48.3	
					83.3
	CTT	Fail	31.2	9.5	
		Pass	8.0	51.4	
					82.6
Vocabulary	IRT Model	Fail	34.9	7.6	
		Pass	8.1	49.4	
					84.3
	CTT	Fail	32.7	8.8	
		Pass	7.8	50.7	
					83.4
Reading	Testlet Model	Fail	34.6	9.8	
		Pass	8.4	47.2	
					81.8
	IRT Model	Fail	35.5	7.7	
		Pass	7.5	49.3	
					84.8
	Poly-IRT Model	Fail	33.6	10.2	
		Pass	9.3	46.8	
					80.4
	CTT	Fail	36.3	8.0	
		Pass	8.0	47.7	
					84.0

advantage of this practice is that a broad content area can be covered within a limited timeframe, which benefits for both validity and reliability. However, one major disadvantage, as demonstrated repeatedly in this article, is that special attention has to be directed to the testlet effect.

Findings from this study support using the testlet response model for language proficiency classification. All investigated tests with testlets violate the local independence assumption and exhibit a strong testlet effect. Although high consistency is observed in the classification results based on the TRT and IRT models, using the IRT model would give test users a wrong impression of how many classification errors may have been committed. From the test design perspective, this could hinder future efforts to improve overall test quality.

This research supports the current practice of conducting a proficiency classification for listening and GCVR tests. These tests are reliable and satisfactory proficiency classification accuracy can be achieved. On the other hand, proficiency classification may not be extended to such subtests as grammar, vocabulary, and reading, without loss of accuracy. Compared to the tests using independent items only (e.g. the vocabulary and

Table 7. Classification consistency between TRT and other models: Listening, GCVR, and Reading Tests

Tests	Measurement models	Classified proficiency levels	TRT Model		Consistency
			Fail	Pass	
Listening	IRT Model	Fail	48.2	0.8	
		Pass	0.8	50.2	
					98.4
	Poly-IRT Model	Fail	47.6	1.4	
		Pass	1.4	49.6	
					97.2
	CTT	Fail	47.0	4.5	
		Pass	2.0	46.4	
					93.4
GCVR	IRT Model	Fail	41.1	1.9	
		Pass	1.9	55.1	
					96.2
	Poly-IRT Model	Fail	41.5	1.5	
		Pass	1.5	55.5	
					97.0
	CTT	Fail	40.8	3.1	
		Pass	2.2	53.9	
					94.7
Reading	IRT Model	Fail	40.6	2.4	
		Pass	2.4	54.6	
					95.2
	Poly-IRT Model	Fail	40.1	3.0	
		Pass	2.4	54.0	
					94.1
	CTT	Fail	38.3	0.8	
		Pass	4.7	56.2	
					94.5

grammar tests), tests with testlets only are more susceptible to both the positive and negative errors. However, when combined with sufficient independent items, the testlet effect could be mitigated and testlet items may pose little threat to proficiency classification, as reflected in the results shown above for the listening and GCVR tests.

Overall, the CTT procedure provides slightly lower classification accuracy than the IRT and TRT procedures. Even for tests where CTT offers comparable results, caution should still be exercised in selecting this model for the proficiency classification. Results from CTT are usually more sample dependent than those from IRT and TRT (Hambleton and Swaminathan, 1985). Moreover, CTT relies on strong assumptions that are hard to meet and test in most test data. This research assumes that a certain percentage of examinees are masters. In practice, standards are usually set up by content experts (Cizek, 2001). In that case, classification errors may be quite different under CTT and TRT in language testing.

Compared to the other three models, it is harder to implement the TRT classification procedure as its proficiency estimation is usually based on the Markov chain Monte Carlo (MCMC) method. This procedure takes a longer time and requires special attention to estimation convergence. However, computer programs such as Scoright and Winbugs (Spiegelhalter et al., 2003) have greatly lessened the technical complexities in applying the TRT model.

This study has investigated language proficiency classification using a pass/fail scheme as implemented for the ECPE test. For situations where the classification decision requires more than two proficiency levels (e.g. advanced, proficient, basic, and below basic), more than one cut-off score should be defined. The measurement models and classification procedures as discussed in this article will remain unchanged. The only difference will be that the expected and observed proportions have to be calculated for more than two categories (Rudner, 2003). While this research has evaluated the accuracy of the proficiency classification based on a fixed cut-off score, it might also be interesting to examine classification accuracy along the proficiency continuum so that classification error could be minimized. For this purpose, the PPoP curve method (Wainer et al., 2005) would be a useful alternative.

The current study has investigated the impact of applying different measurement models to language proficiency classification. The findings have provided some clear guidelines on how proficiency classification can be conducted for language tests. It should nevertheless be noted that procedures studied in this research are more suitable for tests with a large number of items. For tests with a limited number of items (such as the writing and speaking tests of the ECPE), proficiency classification is more challenging, and thus an area to which more future research should be devoted.

Acknowledgements

This research was funded by the University of Michigan, English Language Institute, Spaan Fellowship for Studies in Second or Foreign Language Assessment.

References

Allen MJ and Yen WM (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole.

Adams RJ, Griffin PE and Martin L (1987). A latent trait method for measuring a dimension in second language proficiency. *Language Testing, 4*(1), 9–28.

Alderson JC (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, *13*(2), 219–223.

Bachman LF (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16*(1), 61–70.

Bachman LF (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671–704.

Bachman LF and Palmer AS (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Birnbaum A (1968). Some latent trait models and their use in inferring an examinee's ability. In M. Lord and M. R. Novick (Eds), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Bradlow ET, Wainer H and Wang X (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168.

Brennan RL (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy*, *Version 1.1* (CASMA Research Report No. 9). Iowa City, IA: University of Iowa.

Canale M and Swain M (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Cizek G (Ed). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

English Language Institute, University of Michigan. (2006). *Examination for the Certificate of Proficiency in English 2004–05 annual report*. Ann Arbor, MI: English Language Institute, University of Michigan.

Guo FM (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research and Evaluation*, *11*(*6*).

Keller LA, Swaminathat H and Sireci SG (2003). Evaluating scoring procedures for context dependent item sets1. *Applied Measurement in Education*, *16*(3), 207–222.

Hambleton R and Novick M (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*(3), 159–170.

Hambleton RK and Swaminathan H (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Hanson BA and Brennan RL (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27*(4), 345–359.

Harwell MR and Janosky JE (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*(3), 279–291.

Huynh H (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*(4), 253–264.

Kolen MJ, Hanson BA and Brennan RL (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, *29*, 285–307.

Lee W, Hanson BA and Brennan RL (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*, 412–432.

Livingston SA and Lewis C (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197.

Lord FM (1965). A strong true-score theory with applications. *Psychometrika*, *30*(3), 239–270.

Lord FM (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum.

Lynch BK, Davidson F and Henning G (1988). Person dimensionality in language test validation. *Language Testing, 5*(2), 206–219.

McNamara TF (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52–76.

Nunnally JC and Bernstein IH (1994). *Psychometric theory*. New York: McGraw-Hill.

Oller JW Jr (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23*, 105–118.

Oller JW Jr (1979). *Language tests at school: A pragmatic approach.* London: Longman.

Peterson NS, Kolen MJ and Hoover HD (1989). Scaling, norming, and equating. In Linn, R. L. (Ed) *Educational Measurement* (3rd Edn.) (pp. 221–262). New York: American Council on Education and Macmillan.

Rosenbaum PR (1988). Item bundles. *Psychometrika*, *53*(3), 349–359.

Rudner LM (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research and Evaluation*, *7*(14).

Rudner LM (2003). *The Classification Accuracy of Measurement Decision Theory.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 23–25 April 2003.

Rudner LM (2005). Expected classification accuracy. *Practical Assessment Research and Evaluation, 10*(13). Available online: http://pareonline.net/getvn.asp?v=10andn=13.

Samejima F (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph Supplement, No. 17.

Saito Y (2003). Investigating the construct validity of the cloze section in the Examination for the Certificate of Proficiency in English. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2*, 39–82.

Sireci SG, Thissen D and Wainer H (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247.

Spearman C (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*, 271–295.

Spiegelhalter D, Thomas A and Best N (2003). WinBUGS version 1.4 [computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

Subkoviak M (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*(4), 265–275.

Thissen D (1991). MULTILOG 6.3 [Computer program]. Mooresville, IN: Scientific Software.

Thissen D, Steinberg L and Mooney JA (1989). Trace lines for testlets: A use of multiplecategorical-response models. *Journal of Educational Measurement*, *26*(3), 247–260.

Wainer H and Kiely GL (1987). Item clusters and computerized-adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201.

Wainer H, Wang X, Skorupski WP et al. (2005). A Bayesian method for evaluating passing scores: the PPoP curve. *Journal of Educational Measurement*, *2*(3), 271–281.

Wan L, Brennan RL and Lee W (2007). Estimating classification consistency for complex assessments. (CASMA Research Report No. 22). Iowa City, IA: University of Iowa.

Wang X, Bradlow ET and Wainer H (2002). A general Bayesian model for testlets: Theory and application. *Applied Psychological Measurement*, *26*(1), 109–128.

Wang X, Bradlow ET and Wainer H (2004). *User's guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis.* Research Report 04–49. Princeton, NJ: Educational Testing Services.

Yen WM (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.