



## **Investigating Proficiency Classification for the Examination for the Certificate of Proficiency in English (ECPE)**

**Bo Zhang**

University of Wisconsin–Milwaukee

**ABSTRACT** Accuracy of proficiency classification was evaluated for the Examination for the Certificate of Proficiency in English (ECPE) by comparing procedures based on four measurement models: classical test theory (CTT) model, item response theory (IRT) model, testlet response theory (TRT) model, and polytomous item response theory (Poly-IRT) model. For tests relying on testlets, the TRT model should be chosen for proficiency classification to account for strong testlet effects in examinees' responses; using standard IRT model would inflate classification accuracy. This study shows high accuracy can be achieved for proficiency classification based on the listening and GCVR tests, but not on the cloze and reading tests. The impact of using CTT and Poly-IRT models is also discussed.

Determining whether or not examinees are proficient is one of the most important functions of testing. The value of proficiency classification can be seen from the increasing use of test results for licensure, certification, and selection in many fields. For all these purposes, examinees must be classified into different proficiency levels, such as masters versus nonmasters. By nature, classification decisions are high-risk because any errors may result in misplaced individuals being deprived of well-deserved educational or career development opportunities. Unfortunately, in educational testing, it is almost impossible to avoid measurement errors in estimating proficiency levels. Consequently, classification errors are also unavoidable, and thus accuracy in representing proficiency categories by test scores is of great importance.

Proficiency classification has played a vital role in second language testing. It is one of the major, if not the only, reasons many language learners 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 (MELAB), serve to classify examinees to some degree. In using scores from these tests, universities and colleges usually classify applicants into masters and nonmasters in evaluating language skills.

Language proficiency indicates a person's general communicative competence in the target language environment (Canale & Swain, 1980). Analytically, this proficiency can be broken down into multiple components. From curricular and functional perspectives, it can be

separated into listening, speaking, reading, and writing skills. From a theoretical perspective, Bachman and Palmer (1996) provide a two-tier hierarchical structure of language proficiency. In the first tier is the organizational knowledge, which can be further divided into grammatical and textual knowledge. In the second tier is pragmatic knowledge, which includes lexical, functional, and sociolinguistic competences. This structure prescribes that a proficient language speaker should not only demonstrate structural knowledge but also implement that knowledge effectively in actual use. While language tests are generally assembled by curricular areas, different linguistic competences are also well covered (English Language Institute, 2006).

The above analysis of language proficiency structure leaves practitioners multiple options on how to conduct language proficiency classification. On the one hand, when language proficiency is treated as a general communicative competence, a unidimensional measurement model may be applied to multiple components of a language test and proficiency of each examinee will be represented by a unified score. If, on the other hand, language proficiency is treated as an overarching term encompassing a number of distinct competences, each competency needs to be analyzed separately. Moreover, proficiency classification may be implemented by using models based on either classical test theory or item response theory. For the latter, multiple models could be chosen to analyze one test.

Faced with these options, practitioners need to be informed of the results and consequences in adopting any specific classification procedure. Unfortunately, relevant research on language proficiency classification is extremely limited. While a number of proficiency classification methods have been proposed (e.g., Hanson & Brennan, 1990; Livingston & Lewis, 1995; Rudner, 2001; Wainer, Wang, Skorupski, & Bradlow, 2005), none of them has been carefully studied for language tests. Consequently, it is unclear which procedure creates the fewest classification errors in practice.

The main purpose of this study is to investigate how procedures based on different measurement models affect proficiency classification of ECPE examinees. Two objectives guide this research. The first is to evaluate classification accuracy under four measurement models, which are classical test theory (CTT) model, item response theory (IRT) model, testlet response theory (TRT) model, and polytomous item response theory (Poly-IRT) model. Clearly, the fewer classification errors a model makes, the more valuable it is in practice. The second objective is to study the consistency of classification results when different measurement models are applied. In particular, results from the TRT model are compared to those from the other three models.

## **Competing Measurement Models**

### **Classical Test Theory (CTT)**

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

$$X = T + E \quad (1)$$

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 of an examinee is the average of the observed scores from an infinite number of measurements of this examinee. Unfortunately, this definition will not help obtain the exact value of any true score, thus the true score under CTT remains a theoretical construct.

Test reliability under CTT is defined as:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}, \quad (2)$$

where  $X$ ,  $T$ , and  $E$  have been defined in Equation 1,  $\rho_{XX'}$  is the reliability coefficient,  $\sigma_T^2$  is true score variance, and  $\sigma_X^2$  is observed score variance. To estimate measurement error, standard error of measurement (SEM) is usually computed as:

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

where  $\sigma_X$  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.

Advantages of conducting proficiency classification under CTT are obvious. As shown in Equation 1, the measurement model is simple. Over the relatively long history of educational and psychological testing, methods for estimating test reliability have been well developed for almost all testing situations, such as by split-half reliability (Spearman, 1910), coefficient alpha (Cronbach, 1951), or generalizability theory (Brennan, 2001). These statistics have also been incorporated into popular statistics software packages such as SAS and SPSS. The disadvantages of using CTT, on the other hand, are also apparent and somehow insurmountable. As discussed above, point estimation of the true score is never possible under CTT. As a result, in proficiency classification, it is never known whether an examinee's true score is above or below a cutoff score. Furthermore, CTT assumes a constant  $SEM$  across all examinees. This assumption is hard to meet in most testing conditions. To estimate true proficiency levels with person-specific measurement error, one has to turn to models based on item response theory.

### Item Response Theory (IRT)

Item response theory (Lord, 1980) has gradually developed into the mainstream theory in educational measurement. It is currently applied in most large-scale standardized achievement tests (e.g., SAT, ACT, GRE, LSAT, and MCAT) as well as most state accountability tests. IRT models reflect the interaction between test items and test takers by 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):

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

where  $p$  is the conditional probability that response  $Y_{ij}$  from person  $j$  to item  $i$  is correct,  $\theta$  is the underlying proficiency or ability level,  $c$  is the guessing parameter,  $a$  is the item discrimination parameter,  $b$  is the item difficulty parameter, and  $D$  is a scaling factor. 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

reduces to a two-parameter model. If the discrimination parameter can be further assumed to be constant across all items, a one-parameter IRT model with only the item difficulty parameter may be applied.

Different from CTT in which the true score is known only through a confidence interval, using IRT model will provide direct estimation of theta in the above equation. This theta can be interpreted as the true ability. When the maximum likelihood estimation method is used for tests with large sample sizes, theta estimates are asymptotically unbiased. In addition, a person-specific standard error can also be obtained. Based on these estimates, one can easily classify the proficiency level of each examinee. They can also directly evaluate the likelihood that a positive or negative classification error would be committed.

### Testlet Response Theory (TRT)

An important assumption under the IRT model is local independence (Hambleton & 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, or these responses should be independent. This assumption can also be expressed as that no ability dimension other than the targeted one should affect examinees' 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 & Kiley, 1987). Testlets are commonly employed in language assessments. A classic example 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 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, or these items are locally dependent.

When the local independence assumption is violated, using the standard IRT model would not provide appropriate interpretation of test results as the model no longer fits test data. Specifically, the discrimination parameter would be overestimated (Yen, 1993). As the discriminating power of test items represents how much information an item contributes to ability estimation, overall test information would also be overestimated (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989). Hence, the major harm that local dependence (LD) does to IRT modeling is the inflation of measurement precision.

One direct way to handle LD effect is to model a testlet effect in the IRT model. A testlet response (TRT) model (Bradlow, Wainer, & Wang, 1999) is formulated as:

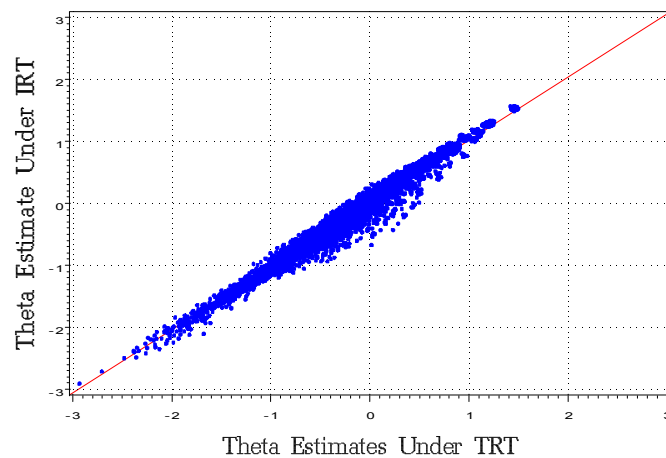
$$P(Y_{ij} = 1 | \theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-Da_i(\theta_j - b_i - \lambda_{id(j)})}}. \quad (5)$$

Compared to the standard IRT model as expressed in Equation 4, the only difference here is the term  $\lambda_{id(j)}$ , which is the testlet effect for person  $j$  in answering item  $i$  nested within testlet  $d(j)$ . The term  $\lambda_{id(j)}$  is assumed to be centered around 0. Its variance indicates the severity of local dependence. If the variance of  $\lambda_{id(j)}$  is zero, there would be no testlet effect and the models in Equations 4 and 5 would be identical. When the variance of  $\lambda_{id(j)}$  is large, testlet

effect will have an impact on the precision of ability estimation (Bradlow, Wainer, & Wang, 1999).

In Figure 1, test information inflation due to LD items is illustrated by the ECPE reading test. In this test, examinees are given 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 from applying the standard 3PL model that totally ignores any possible LD effect. In Figure 1a, not much difference was observed between point estimates of ability by these two models. Most points in the figure are close to the 45-degree reference line, indicating estimates from these two models are about equal. However, in Figure 1b, the standard error of ability estimates from the testlet model is larger than that from the IRT model for most examinees. What this implies is that if the IRT model were selected for proficiency estimation of this reading test, test users would be overconfident about their measurement precision. The IRT model would show greater measurement precision than is, in fact, the case.

1a: Ability Estimate



1b: Standard Error

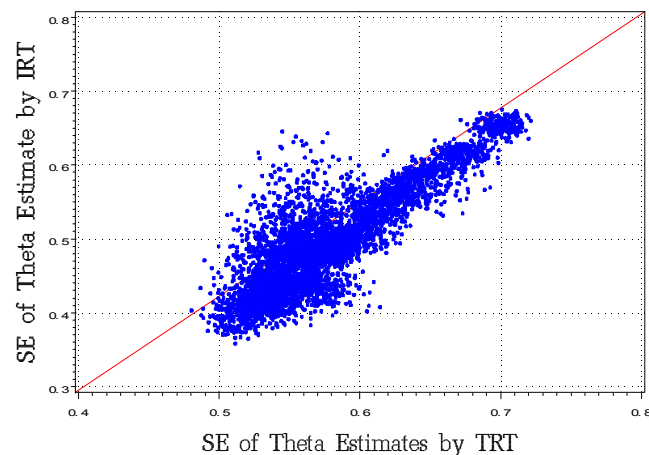


Figure 1. Comparison of Ability Estimates and Standard Error from IRT and TRT

### Polytomous Item Response Models (Poly-IRT)

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

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

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

$$P_{ijk}(\theta) = P_{ijk}^*(\theta) - P_{ij(k+1)}^*(\theta), \quad (7)$$

While the collapsing of locally dependent items may effectively eliminate LD effect, one potential problem is the loss of test information (Yen, 1993). This loss may be quite severe for testlets with a large number of locally dependent items. For example, the cloze test in the present study has 20 items, all based on the reading of one passage. Proficiency estimation using the polytomous IRT model would rely on responses to only one polytomous item instead of those to the original 20 dichotomous items. Using the polytomous model would fail to reflect how examinees have responded to each individual item in the 20-item testlet.

### 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 & Novick, 1973). In educational testing, accuracy must be estimated because errorless test scores never exist. Any misclassified examinee would indicate a classification error. A false positive error occurs when an examinee is classified at a level higher than the true level, whereas a false negative error results when an examinee is put into a category lower than the 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 classification results based on scores from two equivalent forms of the same test. If examinees had been consistently classified into same categories by both forms, classification accuracy would be high. The challenge of this method lies in the difficulty to justify testing same examinees twice by the same test. Accordingly, classification accuracy has to be evaluated based on a single test administration. Over the years, a number of such procedures have been developed, some based on classical test theory (Hanson & Brennan, 1990; Huynh, 1976; Lee, Hanson, & Brennan, 2004; Livingston & Lewis, 1995; Subkoviak, 1976) and others on item response theory (Rudner, 2001; 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 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 to estimate classification accuracy (Brennan, 2004). For the exact steps and technical details of the LL method, refer to Livingston and Lewis (1995).

One particularly useful feature of the LL procedure is that it creates “effective test length” to accommodate both dichotomously and polytomously scored items. Effective test length refers to “the number of discrete, dichotomously scored, locally independent, equally difficulty items required to produce a test score of the same reliability” (Livingston & Lewis, 1995, p. 186). With regard to the effectiveness of the LL procedure, using simulated data, Wan, Brennan, and Lee (2007) found 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 approximation of the true score distribution is unnecessary. The major challenge becomes how to account for measurement error associated with each point estimate of proficiency levels. Rudner (2001, 2005) introduced a method for evaluating decision accuracy through the computation of expected likelihood of classifications. In the following, without loss of generality, this method was described using a pass/fail classification scheme.

Suppose the cut score for passing a test is  $\theta_c$  and the true ability is  $\theta_n$  for Examinee A and  $\theta_m$  for Examinee B. Their positions on the ability scale are depicted in Figure 2. Due to the error associated with ability estimation, a conditional distribution accompanies each theta estimate. As  $\theta_n$  is smaller than  $\theta_c$ , Examinee A should be classified as a nonmaster. Accordingly, 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 theta estimates will be larger than  $\theta_c$ . In classification terminology, this chance is the likelihood that a false positive error has been committed as a true nonmaster would be identified as a master.

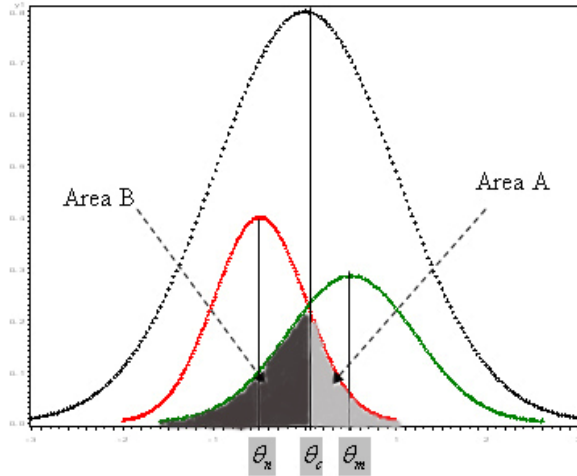


Figure 2. Illustration of False Positive and Negative Errors in Proficiency Classification

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

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

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

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

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

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

where  $M$  is the number of masters. The expected frequencies for correct classifications, either true nonmasters as nonmasters or true masters as masters, could be computed in the same manner as in Equations 9 and 10. These expected frequencies will then be compared to the observed frequencies to evaluate classification accuracy.

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



## Methods

### Instruments and Subjects

The Examination for the Certificate of Proficiency in English (ECPE) is an English language proficiency test for adult nonnative 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 the proficiency classification for two sections in the ECPE: listening and GCVR. Both sections have a large number of items, allowing investigation 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 one prompt. The last 15 items are based on listening to three long dialogues or paragraphs, each followed by five questions. Because these items are locally dependent, they are susceptible to testlet effect. In the cloze test, all 20 items share one stimulus. Thus it is possible that there is strong testlet effect. The grammar and vocabulary tests each use 30 independent items. The reading test asks students to read four paragraphs, each followed by five questions. Local dependence may exist in this test as well. All test items studied in this research are multiple-choice items.

Subjects for this study were 5,000 examinees, randomly selected from a one-year administration of the ECPE.

### Measurement Models and Model Estimation

Table 1 lists the measurement models that were applied to various tests. Note that proficiency classification was conducted by using the listening and GCVR tests in practice. In studying the GCVR test, it was assumed that there was one general language competence measured by all items. Next, four subtests under GCVR were examined separately, assuming they measured distinct language competences. Results from this analysis would provide information on whether accuracy could be achieved if classification were to be conducted at these subtest levels.

Table 1. Measurement Models Applied for Proficiency Classification

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

The grammar and vocabulary tests use no testlet, thus results from TRT would be equivalent to those from IRT. The listening, cloze, reading, and GCVR tests all showed strong testlet effect. The appropriate model would be the TRT model. The IRT model was applied to study possible damage to proficiency classification if local dependence is ignored.

As all items were multiple-choice items, a three-parameter logistic model was applied for both the IRT and TRT models to account for a possible guessing effect. Proficiency estimation under the IRT model was obtained by using the MULTILOG computer program (Thissen, 1991). This program implemented the marginal maximum likelihood method to estimate the ability trait. To increase 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 is large (i.e., 5,000), the impact of these priors on  $a$ 's and  $b$ 's was probably quite limited (Harwell & Janosky, 1991). The main purpose of using these priors was to constrain the  $c$  parameter to reasonable values. Proficiency estimation for the Poly-IRT model was also conducted by using the MULTILOG program. For the cloze test, when all 20 items were collapsed into one 20-category polytomous item, convergence could not be achieved in proficiency estimation. As a compromise, these items were collapsed into two 10-category items.

For the TRT model, parameter estimation was based on the Markov chain Monte Carlo procedure, as implemented in the Scorigt program (Wang, Wainer, & Bradlow, 2004). This program adopts a full Bayesian hierarchical structure to estimate testlet model parameters. For details on the estimation algorithm, refer to Wang, Bradlow, and Wainer (2002), and the Scorigt program manual (Wang, Bradlow, & Wainer, 2004). One important issue in the MCMC estimation is convergence of the posterior distribution of each model parameter to a stationary distribution. Following the suggestions in the Scorigt manual on how to improve and check 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. While convergence was achieved for all tests, it is noteworthy that chains had to run considerably longer (25,000 iterations) for the cloze test than for other tests.

### **Cut Scores for Proficiency Classification**

Cut scores were identified for classifying examinees into two proficiency categories: pass or fail. The original cut score for passing the listening test was 34 on the raw score scale, which corresponded to 48 on the percentile rank scale<sup>1</sup>. It was thus assumed that 52% of the examinees from this test administration were masters. This assumption helped to set up the cut scores for the IRT and TRT models. Specifically, values on the  $\theta$  scale that corresponded to the percentile rank of 52 were used as cutoff scores. For the GCVR test, the cut score and the percentile rank were 48 and 42, respectively.

---

<sup>1</sup> Editor note. This raw score is an approximation. The ECPE tests use IRT theta estimation to set pass levels.

## Results

Before any proficiency classification was conducted, how well test items had measured relevant ability traits was evaluated. This is important because items not measuring the corresponding trait properly might invalidate the application of measurement models. For that purpose, the corrected point-biserial correlation between item responses and the 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 indicates these items had a negative discrimination parameter. Examinees with higher ability would be less likely to answer these items correctly. Consequently, these items were not used for proficiency classification.

Table 2 presents reliability estimates for tests under study. Using 0.7 as the general guideline of acceptable reliability level (Nunnally & Bernstein, 1994), reliability was high enough for all but the cloze test. The listening and GCVR tests show the highest reliability, which certainly is important as classification results are based on these section scores in practice. The cloze test, on the other hand, had the lowest reliability, partially due to its shortest length among all the tests.

Table 2. Test Reliability Coefficient

Subtests	No. of Items	Reliability
Listening	48	.76
Grammar	30	.70
Cloze	20	.58
Vocabulary	29	.73
Reading	20	.74
GCVR	99	.86

Next, testlet effect was examined. Following the suggestion by Bradlow, Wainer, and Wang (1999), a variance over .3 for the testlet term  $\lambda_{id(j)}$  was set as the criteria to decide if testlet effect is significant. Table 3 shows all tests using testlets had strong testlet effect, which confirmed that reading and listening passages in the ECPE test violated the local independence assumption. As expected, the cloze test demonstrated the strongest testlet effect. Note that the same testlets, when placed in different tests, exhibited different magnitudes of 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 reading items became stronger in the GCVR test than in the reading test. While it is hard to illuminate this testlet effect change by studying item responses only, analyzing the content of these testlets may shed some light on these shifts.

Table 3. Magnitude of the Testlet Effect

Subtest	No. of Testlets	Testlet Effect
Listening	1	1.05
	2	.41
	3	1.09
Cloze	1	1.43
Reading	1	.58
	2	.53
	3	.35
	4	.59
GCVR	1	.27
	2	1.31
	3	1.44
	4	1.33
	5	1.83

Table 4 presents classification results for the listening and GCVR tests. Table cell values indicate the percentage of examinees falling into each category. The accuracy column gives the percentage of examinees correctly classified. Use the listening test as an example. Under the testlet model, 41.5% of examinees had been correctly identified as not passing and 43.3% as passing. To add up these two numbers, 84.8% of total examinees were accurately identified. Meanwhile, 7.5% of examinees would be misclassified as masters. They were positive errors. Negative errors are those 7.7% examinees who were expected to pass but were actually classified as failing.

Overall, classification accuracy was impressively high for both tests. Results based on different measurement models were quite similar. For the listening test, about 85% of examinees were expected to be correctly identified. That is to say about 15% of examinees could be misclassified. Classification accuracy was slightly higher for the GCVR test with the percentage of agreement around 87%. For both the listening and GCVR tests, the false positive error rate and the false negative error rate were about equal.

Table 4. 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	84.8	
		Pass	7.5	43.3		
	IRT Model	Fail	41.5	7.6	85.0	
		Pass	7.5	43.5		
	Poly-IRT Model	Fail	41.3	8.0	84.3	
		Pass	7.7	43.0		
	CTT	Fail	39.8	8.1	84.1	
		Pass	7.7	44.3		
	GCVR	Testlet Model	Fail	37.0	5.8	88.1
			Pass	6.0	51.1	
		IRT Model	Fail	37.1	5.6	88.5
			Pass	5.9	51.4	
Poly-IRT Model		Fail	36.8	5.9	87.9	
		Pass	6.2	51.1		
CTT		Fail	35.8	6.0	88.4	
		Pass	5.6	52.6		

Table 5 gives results for the subtests under GCVR. Classification accuracy was clearly lower for these tests. Both false positive and false negative errors increased considerably under all models. Errors were especially higher for the cloze and reading tests, both of which relied heavily on testlets. For the cloze test, lower than 70% of the examinees were expected to be correctly classified under the TRT model and classification errors were also high under other models.

Table 5. Classification Accuracy: Grammar, Cloze, Vocabulary, and Reading Tests

Tests	Measurement Models	Classified Proficiency Levels	Expected Proficiency Levels		Accuracy		
			Fail	Pass			
Grammar	IRT Model	Fail	35.0	8.7	83.3		
		Pass	8.0	48.3			
	CTT	Fail	31.2	9.5			
		Pass	8.0	51.4			
						82.6	
	Cloze	Testlet Model	Fail	28.3		16.6	68.7
Pass			14.7	40.4			
IRT Model		Fail	33.5	9.2			
		Pass	9.5	47.8			
Poly-IRT Model		Fail	33.2	11.4	81.3		
		Pass	9.8	45.6			
					78.8		
CTT		Fail	31.8	10.1			
		Pass	11.2	46.9			
					78.7		
Vocabulary		IRT Model	Fail	34.9	7.6	84.3	
			Pass	8.1	49.4		
	CTT	Fail	32.7	8.8			
		Pass	7.8	50.7			
							83.4
	Reading	Testlet Model	Fail	34.6	9.8		81.8
Pass			8.4	47.2			
IRT Model		Fail	35.5	7.7			
		Pass	7.5	49.3			
Poly-IRT Model		Fail	33.6	10.2	84.8		
		Pass	9.3	46.8			
					80.4		
CTT		Fail	36.3	8.0			
		Pass	8.0	47.7			
					84.0		

For the cloze and reading tests, there were fewer classification errors under the IRT than under the TRT. This higher accuracy should not be interpreted as the IRT model providing more accurate proficiency classification. As shown in Table 3, both the cloze and

reading tests had strong testlet effect, thus 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, consistency of classification between the TRT model and the other three models are reported. Results for the listening and GCVR tests are given in Table 6 and those for GCVR subtests are in Table 7. For all tests, the agreement between the TRT and IRT classifications was above 95%. This is even true for tests with strong testlet effect. The high constancy observed in this table was expected as point estimates of proficiency level were very similar under these two models, as exemplified by the reading test in Figure 1.

Table 6. Classification Consistency between TRT and Other Models: Listening and GCVR Tests

Tests	Measurement Models	Classified Proficiency Levels	TRT Model		Consistency
			Fail	Pass	
Listening	IRT Model	Fail	48.2	0.8	98.4
		Pass	0.8	50.2	
	Poly-IRT Model	Fail	47.6	1.4	97.2
		Pass	1.4	49.6	
	CTT	Fail	47.0	4.5	93.4
		Pass	2.0	46.4	
GCVR	IRT Model	Fail	41.1	1.9	96.2
		Pass	1.9	55.1	
	Poly-IRT Model	Fail	41.5	1.5	97.0
		Pass	1.5	55.5	
	CTT	Fail	40.8	3.1	94.7
		Pass	2.2	53.9	

The Poly-IRT model provided comparable results to the TRT model for the listening and GCVR tests. The discrepancy between these two models was considerably larger for the reading and cloze tests, likely due to the fact that all items in these two tests were collapsed into polytomous items. A significant amount of test information might have been lost. For the listening and GCVR tests, the CTT model tended to classify more masters under TRT as nonmasters than vice versa. But for the cloze and reading tests, the opposite was true. For example, in the cloze test, as many as 8% of masters under TRT would be classified as nonmasters under CTT.

Table 7. Classification Consistency between TRT and Other Models: Grammar, Cloze, Vocabulary, and Reading Tests

Tests	Measurement Models	Classified Proficiency Levels	TRT Model		Consistency
			Fail	Pass	
Cloze	IRT Model	Fail	42.6	0.4	99.2
		Pass	0.4	56.6	
	Poly-IRT Model	Fail	39.4	3.5	92.9
		Pass	3.6	53.5	
	CTT	Fail	35.0	2.1	89.9
		Pass	8.0	54.9	
Reading	IRT Model	Fail	40.6	2.4	95.2
		Pass	2.4	54.6	
	Poly-IRT Model	Fail	40.1	3.0	94.1
		Pass	2.4	54.0	
	CTT	Fail	38.3	0.8	94.5
		Pass	4.7	56.2	

### Discussion

The Examination for the Certificate of Proficiency in English (ECPE) is a representative example of the commonly used second language tests. These tests measure general language competence by assessing skills in listening, reading, grammar, vocabulary, and writing activities. They routinely use long passages as prompts to ask a large number of multiple-choice items. The main advantage of this practice is a broad content area can be covered in a limited amount of testing time, which helps increase both test reliability and test validity. However, one major disadvantage of this practice, as demonstrated repeatedly in this study, is that special attention has to be paid to testlet effect.

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

This research also supports the current practice of conducting proficiency classification based on the listening and GCVR tests. These tests are highly reliable and high proficiency classification accuracy could be achieved. On the other hand, proficiency classification could not be extended to the GCVR subtest level as classification accuracy would be considerably lower. For example, the cloze test would pose great challenges if



studied separately. No matter which measurement model one chooses, it is likely that classification error rates would be high. To achieve high accuracy, this test almost has to be combined with other tests. In addition, while cloze test items may have linguistic advantages over other type of items in measuring language competence (Sato, 2003), experience with the present ECPE data indicates that it is not easy to obtain stable estimation of language proficiency based on the cloze test items alone.

Testlets show a negative impact on proficiency classification. Compared to tests using independent items only, testlet items increase measurement error, which in turn cause classification more susceptible to both false positive and false negative errors. This impact is especially serious for tests with all items in testlets, such as the cloze and reading tests in this study. However, when combined with sufficient independent items, testlet items may pose little threat to proficiency classification, as observed in the studied listening and GCVR tests.

For tests where CTT seems able to offer results comparable to those from TRT or IRT, caution should still be exercised in selecting the CTT model for proficiency classification. As discussed in the section of “Competing Measurement Models,” CTT relies on strong assumptions that are hard to meet in most test data. In addition, results from this study may be limited by how the cut scores have been decided. This research assumes that a certain percentage of examinees were masters. In practice, standards are usually set up by content experts (Cizek, 2003). In that case, classification errors may be quite different under CTT and TRT models in language testing.

Compared to the other three models, it is harder to implement the TRT model for proficiency classification as proficiency estimation is generally based on the Markov chain Monte Carlo procedure. This procedure typically takes a long time and requires special attention to estimation convergence. However, computer programs such as Scoright and Winbugs (Spiegelhalter, Thomas, & Best, 2003) have greatly reduced technical complexities in applying the TRT model.

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

## References

- Bachman, L. F., & Palmer, A. S. (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, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64(2), 153–168.
- Brennan, R. L. (2001). *Generalizability Theory*. New York : Springer-Verlag.

- Brennan, R. L. (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: University of Iowa.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.
- Cizek, G. (2001). (Ed.). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, N.J.: Erlbaum.
- Cronbach, L. J. (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: English Language Institute, University of Michigan.
- Guo, F. M. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation*, *11*(6).
- Keller, L.A., Swaminathat, H., & Sireci, S.G. (2003). Evaluating scoring procedures for context dependent item sets1. *Applied Measurement in Education*, *16*(3), 207–222.
- Hambleton, R., & Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, *10*(3), 159–170.
- Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.
- Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27*(4), 345–359.
- Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*(x), 279–291.
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*(4), 253–264.
- Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*(x), 412–432.
- Liu, Y., Bolt, D.M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *1*(1), 3–21.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197.
- Lord, F.M. (1965). A strong true-score theory with applications. *Psychometrika*, *30*(3), 239–270.
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates, Inc.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.
- Rosenbaum, P.R. (1988). Item bundles. *Psychometrika*, *53*(3), 349–359.
- Rudner, L.M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, *7*(14).
- Rudner, L.M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation*, *10*(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. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 2, 39–82.
- Sireci, S. G., Thissen, D., & 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. & 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. & Mooney, J.A. (1989). Trace lines for testlets: A use of multiplecategorical-response models. *Journal of Educational Measurement*, 26(3), 247–260.
- Wainer, H., Bradlow, E.T., & Du, Z. (2000). Testlet response theory: an analog for the 3- PL useful in testlet-based adaptive testing. In W.J. van der Linder & C.A.W. Glas (Eds.), *Computerized adaptive testing: theory and practice*, 245–270. Boston, MA: Kluwer-Nijhoff.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory*. New York: Cambridge University Press.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized-adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24(3), 185–201.
- Wainer, H., Wang, X., Skorupski, W. P., & Bradlow, E. T., (2005). A Bayesian method for evaluating passing scores: the PPOP curve. *Journal of Educational Measurement*, 2(3), 271–281.
- Wan, L., Brennan, R. L., & Lee, W. (2007). Estimating Classification Consistency for Complex Assessments. (CASMA Research Report No. 22). Iowa City, IA: University of Iowa.
- Wang, W. C. & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, 29(4), 296–318.
- Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: theory and application. *Applied Psychological Measurement*, (26) No. 1, 109–128.
- Wang, X., Bradlow, E. T., & 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, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.