International Journal of Education and Psychology in the Community
IJEPC
2016, 6 (1 & 2), July, 106-120

THE INFLUENCES OF INTER-ITEM CORRELATIONS AND SAMPLE SIZES ON THE CLASSIFICATION INDICES UNDER ITEM RESPONSE THEORY: THE SIMULATION STUDY

Narissara Suaklay • **Nuttaporn Lawthong** **Sirichai Kanjanawasee**
Chulalongkorn University, Bangkok, Thailand

Abstract
This study purposed 1) to investigate the influences of inter-item correlations and sample sizes on the classification indices and 2) to compare the classification indices estimation methods under Item Response Theory, which was a simulation study. The data used in the study were secondary ones from Set A of the Ordinary National Education Test: O – NET with 30 items in Mathematic Subject for ninth grade students in the academic year of 2013. The population was 340,084 students. The samples used for initial data analysis were 5,000 units selected with the systematic random sampling in the simulation study from the R program. The results revealed that: firstly, it was found that Lee's method provided insignificantly higher classification accuracy and consistency than Rudner's method after considering cross points and overall images. In the case that the items had low consistency, both estimation methods have non-significantly different accuracies and consistencies at the significance level of .05. In the case that the items had high correlation, there was a chance that accuracies and consistencies might be significantly different at the significance level of .05. If the sample sizes were increased, the two methods might provide significant differences at the significance level of .05. Secondly, if the sample sizes were different regardless of the independencies between the items, it was found that the two methods could estimate the values of the indices with indifferent accuracies and consistencies at the

Correspondence concerning this paper should be addressed to:

• Ph.D. candidate in Educational Measurement and Evaluation, Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University. Bangkok 10330, Thailand E-mail address: *Narissara.Sua@student.chula.ac.th*

significance level of .05. After analyzing the estimation methods, it was found that sample sizes did not influence both methods.

Keywords: classification indices; classification consistency; classification accuracy; item response theory; IRT

Introduction

The purposes of the large-scale testing are to obtain reliable scores and to reasonably translate statements in order to make decisions on placement, remediation, and certification (Wainer & Kiely, 1987). The mentioned decisions must consider the scores from the test in order to compare the scores with the test criteria and cut points. By considering only the test or observed scores, it might lead to classification errors. There are two cases of the classification errors. The first case occurs in two parallel managerial situations. The test administrator must classify or group the examinees into performance groups by considering their test scores. There are two patterns of classification errors. Firstly, the examinees were classified into the higher performance group than standards in any test and then the lower one in the second test.

The possibilities of these two patterns are called as inconsistent classifications. If the examinees from the two situations are classified into the same performance group that is either higher or lower than the standards, this is the consistent classification showing that there is no classification error.

The second case is the classification error occurring in any general managerial situation. The test managers must make decisions to classify the examinees into a particular performance group by considering the test or observed scores as well as the actual conditions of the examines regarding the pattern average or actual scores. There are two patterns of classification errors. Firstly, the examinees are classified into the higher performance group than standards after considering the observed scores. Secondly, they are classified into the lower one after considering the actual scores. The possibilities of these two patterns are called as inconsistent misclassifications. If the examinees from the two situations are classified into the same performance group that is either higher or lower than the standards, this is the correct classification showing that there is no classification error.

The classification errors made the psychological estimators acknowledge the importance of this problem since they do not know the error size of each examinee in any test situation (Wainer & Kiely, 1987). It was also found that the traditional validation estimation developed and based on test scores might not be appropriate for the inconsistency and correctness estimations. Hence, a new and appropriate inconsistency evaluation technique must be used in order to minimize classification errors with the classification indices including classification consistency and accuracy that have been simultaneously developed since 1970.

During the first development period, values were estimated according to the traditional test theory. Particularly, the classification consistency estimation must be done under two managerial test situations since the consistency estimation is the comparison between the examinee classifications in two parallel situations (Huynh, 1976; Subkoviak, 1976; Livingston & Lewis, 1995; Lee, 2007; Lee, Brennan, & Wan, 2009). However, it is difficult to actually conduct two parallel tests. Hence, there were some estimators developing the single administration for classification indices under the item response theory or IRT (Huynh, 1990; Rudner, 2005; Guo, 2006; Lee, 2010).

By reviewing the studies relevant to academic achievement measurements, it could be observed that the decision to use any estimation method for checking different consistencies and accuracies must consider their performances under applied situations in order to efficiently estimate the classification consistencies and accuracies and to minimize errors.

Therefore, the comparisons the indices from different concepts and theories in order to measure the performances of the index evaluation is important in order to obtain information useful for applications in various test applications. This is especially important for Ordinary National Educational Testing (O-NET), which is a high stakes test with evaluation standards.

By considering the classification index evaluation concepts and theories (i.e. the methods of Rudner, 2005; Lee, 2010) under IRT, it was found that the two estimation methods have initial agreements about the IRT model. Nevertheless, the methods have different error distributions for the actual estimations, data characteristics, and probability formulas for classifying the examinees into performance levels. Rudner's method initial specifies that the actual errors must be normal distributions, while Lee's method initially set that it must be the compound binomial distribution regarding data characteristics or

scores. A difference is that Rudner's method use the scores on the theta scale, but Lee's method use the scores on the summed score scale. Regarding the probability formulas for classifying the examinees into performance levels, the methods are different because the agreements on the error distribution of the actual estimation as previously described. With these differences, the researcher chose to study the performance of the two indices estimation methods under IRT.

With the aforementioned reasons, the researcher was interested in investigating the influence of inter-item correlations and sample sizes on the classification indices under IRT in order to compare the performances of the two methods in the simulation study in order to obtain information about the classification indices estimations and to obtain the highest performances under the mentioned situations or conditions useful for selecting indices estimation methods with appropriate consistency and accuracy for the high return situations in the future.

Research objectives

1) To investigate the influences of the inter-item correlations and the sample sizes on the classification indices.

2) To compare the classification indices estimation methods under Item Response Theory (Rudner's methods and Lee's methods).

Relevant Theories

Classification indices

The consistency index is the inconsistency classification that occurs in two parallel managerial situations. The test administrator must classify or group the examinees into performance groups by considering their test scores. There are two patterns of classification errors. Firstly, the examinees were classified into the higher performance group than standards in any test and then the lower one in the second test. The possibilities of these two patterns are called as inconsistent classifications. If the examinees from the two situations are classified into the same performance group that is either higher or lower than the standards, this is the consistent classification showing that there is no classification error.

The accuracy index is the classification error occurring in any general managerial situation. The test managers must make decisions to classify the examinees into a particular performance group by considering the test or observed scores as well as the actual conditions of the examines regarding the pattern average or actual scores. There are two patterns of classification errors. Firstly, the examinees are classified into the higher performance group than standards after considering the observed scores. Secondly, they are classified into the lower one after considering the actual scores. The possibilities of these two patterns are called as inconsistent misclassifications. If the examinees from the two situations are classified into the same performance group that is either higher or lower than the standards, this is the correct classification showing that there is no classification error.

Rudner's Method

The method was developed by Rudner (2005) has initial agreements on the estimation errors for actual scores. This method is usable for both dichotomous and polytomous items as well as those with IRT pattern scores. It also uses the test scored on theta scale as the model for the index estimation.

Lee's Method

It was developed by Lee (2010) with the initial agreements about the compound binomial distribution for estimating actual values. It is the method that can be used both dichotomous and polytomous items as well as those with summed scores. It also uses the mixture IRT model for the index estimation.

Conceptual Framework

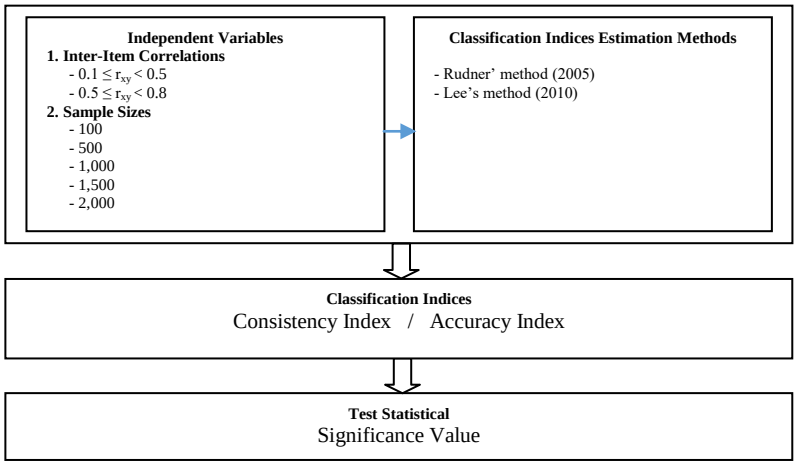

Figure 1. Conceptual Framework

Methods

Participants

Data Sources

The data used in this study were secondary ones from the Ordinary National Educational Testing (O-NET) in Mathematic Subject for ninth grade students in the academic year of 2013. There were two standard sets of answers (i.e. Set A and Set B). Each set had 30 items. The tests were conducted by the National Institute of Educational Testing Service; NIETS (Public Organization) in Thailand. The population included 340,084 people. The samples were 5,000 units from Set A in Mathematic Subject for ninth grade students those selected with the systematic random sampling by using R program.

Data Simulation

Using R program for simulated data from the empirical data above. These data demonstration process contained the differential conditions of independent variables (the inter-item correlations and the sample sizes), those affected classification indices by using the estimation classification methods of Rudner and Lee as shown in Table 1.

Table 1. Data Demonstration Conditions

Process	Data	Demonstration Conditions	
		inter-item correlations	sample sizes
1	Empirical data	$r_{xy} < 0.1$	2000 – 5000*
2	Simulation data	$0.1 \leq r_{xy} < 0.5$	100
3	Simulation data	$0.1 \leq r_{xy} < 0.5$	500
4	Simulation data	$0.1 \leq r_{xy} < 0.5$	1000
5	Simulation data	$0.1 \leq r_{xy} < 0.5$	1500
6	Simulation data	$0.1 \leq r_{xy} < 0.5$	2000
7	Simulation data	$0.5 \leq r_{xy} < 0.8$	100
8	Simulation data	$0.5 \leq r_{xy} < 0.8$	500
9	Simulation data	$0.5 \leq r_{xy} < 0.8$	1000
10	Simulation data	$0.5 \leq r_{xy} < 0.8$	1500
11	Simulation data	$0.5 \leq r_{xy} < 0.8$	2000

The random variable is this study the score (0 or 1) as shown in Table 1. Each situation is randomized for 100 times. The scores were analyzed and compared with the means of accuracies and consistencies.

Procedures

The procedures for this simulation study consisted as follow steps:

1) Collecting empirical data from the Ordinary National Educational Testing (O-NET) in Mathematic Subject for ninth grade students in the academic year of 2013.

2) Estimating item parameters with R package 'psych' and package 'ltm' and analyzing descriptive statistic of empirical data with SPSS program.

3) Simulating the items respond data with R package based on the item parameter estimated from empirical data those were applying to the step of estimating classification indices. The simulated data were demonstrated with differential conditions of independent variables (the inter-item correlations and the sample sizes), those affected classification indices by using the estimation classification methods of Rudner and Lee.

4) Estimating theta score scale cut point for Rudner's method from the seven cut points used in this simulation study were those specified by NIETS with IRTPRO program. This cut points could be used with the method of Lee (2010). For the method of Rudner (2005), the cut points must meet the theta scale. Hence, the raw score scale must be converted into the theta score scale with Test Character Curve (TCC) (Rupp, Templin, & Henson, 2010; Houts & Cai, 2013) as shown in Table 2.

Table 2. The seven cut points in raw score scale and theta score scale

Place of cut score	**raw score scale**	**theta score scale**
1	7.50	0.03
2	9.36	0.49
3	11.04	0.75
4	13.20	1.01
5	16.08	1.29
6	19.92	1.63
7	24.00	1.96

5) Estimating classification indices by using Rudner's method and Lee's method with R package 'cacITR' (Lathrop, 2015) for empirical data and simulation data. The input data in this step were the theta parameter (θ), the standard error of true score estimation (se), the item parameters (difficulty; b_i, discrimination; a_i and guessing; c_i)

6) Comparing the classification indices estimation methods under Item Response Theory (Rudner's methods and Lee's methods) with nonparametric test statistic in two conditions as follow:

6.1) condition with difference of two inter-item correlations ($0.1 \leq r_{xy} < 0.5$ and $0.5 \leq r_{xy} < 0.8$);

6.2) condition with difference of five sample sizes (n=100, 500, 1,000, 1,500 and 2,000).

Results and Interpretation

This study was a simulation study using R program for simulated data and estimated classification indices. The empirical data was the scores from set A of O-NET in Mathematic Subject for ninth grade students in the academic year of 2013. The analysis results were as follows.

1. Empirical Data Analysis Results

1.1 Descriptive Statistical Analysis Results

It was found that the mean score, the standard deviation, the highest score and the lowest score of the 340,084 examinees were 25.27, 11.29, 100 and 0 respectively. By considering the skewness (Sk), it was found that its distribution was right skewness (the skewness was positive). This meant that most examinees had lower scores than the mean score. By considering the kurtosis (Ku), it was found that the score was low distribution (the kurtosis was positive). The Cronbach's alpha coefficient of internal consistency reliability was 0.57.

1.2 Parameter Estimation Results

Analyzing the item response theory model with R program. Estimating item parameters by using maximum likelihood, there were found as follows: the difficulty parameters (b_i) were -3.213 to 3.447, the discrimination parameters (a_i) were -3.757 to 3.964 and the guessing parameters (c_i) were 0 to 0.386. Most items had appropriate difficulties, discriminations and guessing rates.

1.3 Correlation Coefficient Analysis Results

By analyzing the correlation coefficients between the items, there were found that most correlation coefficients were positive at low to middle levels.

The values were -0.009 to 0.218. There were 11 correlation coefficients those were not significantly different at the significance levels of .01 and .05.

2. Classification Indices Estimation Results

2.1 Classification Indices Estimated from the Empirical Data

Considering each cut score and simultaneous were found that Lee's method had higher accuracy and consistency than Rudner's method as shown in Table 3.

Table 3. Accuracy and Consistency Indices Estimated from the Empirical Data

Place of cut score	Rudner		Lee	
	Accuracy	**Consistency**	**Accuracy**	**Consistency**
1	0.8148	0.7388	0.7273	0.6406
2	0.8395	0.7654	0.8572	0.7846
3	0.8543	0.7856	0.9378	0.9059
4	0.8745	0.8135	0.9736	0.9599
5	0.9171	0.8676	0.9893	0.9847
6	0.9590	0.9311	0.9929	0.9903
7	0.9789	0.9638	0.9956	0.9940
Simultaneous	**0.5603**	**0.4576**	**0.5769**	**0.4671**

2.2 Classification Indices Estimated from the Simulation Data

1) The inter-item correlations of $0.1 \leq r_{xy} < 0.5$

In consideration of the inter-item correlations were more than or equal to 0.1 but less than 0.5 as shown in Table 4. According to four sample sizes conditions, they were found that Rudner's method had higher accuracy and consistency than Lee's method all cases in simultaneous, whereas considering each cut score were not found that.

Table 4. Accuracy and Consistency Indices Estimated from the Simulation Data with the inter-item correlations of $0.1 \leq r_{xy}<0.5$

Place of cut score	indices	n							
		100		500		1000		2000	
		Rudner	Lee	Rudner	Lee	Rudner	Lee	Rudner	Lee
1	accuracy	0.90	0.84	0.90	0.92	0.90	0.99	0.90	0.93
	consistency	0.87	0.78	0.87	0.89	0.86	0.98	0.86	0.90
2	accuracy	0.93	0.89	0.92	0.91	0.91	0.97	0.91	0.92
	consistency	0.91	0.86	0.88	0.87	0.88	0.96	0.88	0.89
3	accuracy	0.93	0.94	0.93	0.91	0.93	0.96	0.92	0.91
	consistency	0.91	0.92	0.90	0.88	0.90	0.95	0.89	0.88
4	accuracy	0.96	0.97	0.94	0.93	0.94	0.95	0.94	0.91
	consistency	0.94	0.95	0.92	0.90	0.92	0.93	0.91	0.88
5	accuracy	0.96	0.98	0.96	0.94	0.95	0.93	0.95	0.92
	consistency	0.95	0.98	0.94	0.92	0.94	0.90	0.94	0.90
6	accuracy	0.98	0.99	0.97	0.96	0.96	0.91	0.97	0.95
	consistency	0.97	0.98	0.96	0.94	0.95	0.87	0.96	0.93
7	accuracy	0.99	0.99	0.98	0.98	0.97	0.89	0.98	0.97
	consistency	0.98	0.99	0.98	0.97	0.97	0.86	0.98	0.96
Simultaneous	**accuracy**	**0.71**	**0.68**	**0.67**	**0.62**	**0.67**	**0.65**	**0.66**	**0.61**
	consistency	**0.63**	**0.61**	**0.60**	**0.54**	**0.59**	**0.57**	**0.58**	**0.52**

2) The inter-item correlations of $0.5 \leq r_{xy}<0.8$

In consideration of the inter-item correlations were more than or equal to 0.5 but less than 0.8 as shown in Table 5. According to four sample sizes conditions, they were found that Rudner's method had higher accuracy and consistency than Lee's method all cases in simultaneous and each cut score.

Table 5. Accuracy and Consistency Indices Estimated from the Simulation Data with the inter-item correlations of $0.5 \leq r_{xy}<0.8$

Place of cut score	indices	n							
		100		500		1000		2000	
		Rudner	Lee	Rudner	Lee	Rudner	Lee	Rudner	Lee
1	accuracy	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.94

Table 5. Accuracy and Consistency Indices Estimated from the Simulation Data with the inter-item correlations of $0.5 \leq r_{xy} < 0.8$ - *continued*

Place of cut score		**n**							
	indices	**100**		**500**		**1000**		**2000**	
		Rudner	**Lee**	**Rudner**	**Lee**	**Rudner**	**Lee**	**Rudner**	**Lee**
2	consistency	0.94	0.94	0.93	0.93	0.94	0.93	0.94	0.92
	accuracy	0.96	0.95	0.95	0.95	0.95	0.95	0.96	0.94
3	consistency	0.95	0.94	0.93	0.93	0.93	0.93	0.94	0.92
	accuracy	0.94	0.95	0.95	0.95	0.96	0.95	0.96	0.95
4	consistency	0.92	0.93	0.93	0.94	0.94	0.93	0.95	0.93
	accuracy	0.96	0.94	0.97	0.96	0.97	0.95	0.98	0.95
5	consistency	0.94	0.92	0.96	0.94	0.97	0.94	0.97	0.93
	accuracy	0.97	0.94	0.98	0.96	0.99	0.96	0.99	0.96
6	consistency	0.96	0.93	0.98	0.95	0.98	0.95	0.99	0.94
	accuracy	0.98	0.94	0.99	0.97	0.99	0.96	0.99	0.96
7	consistency	0.98	0.93	0.99	0.95	0.99	0.95	0.99	0.95
	accuracy	0.99	0.93	0.99	0.96	0.99	0.96	0.99	0.96
	consistency	0.99	0.92	0.99	0.95	0.99	0.95	0.99	0.95
Simultaneous	**accuracy**	**0.81**	**0.67**	**0.83**	**0.76**	**0.86**	**0.74**	**0.87**	**0.73**
	consistency	**0.76**	**0.60**	**0.79**	**0.71**	**0.82**	**0.68**	**0.84**	**0.66**

3. The Comparison Results of the Classification Indices Estimation Methods of Rudner and Lee in the Different Demonstration Situations

In consideration of the low inter-item correlations ($0.1 \leq r < 0.5$) were found that the classification indices estimated by Rudner's method non-significantly difference with Lee's method, but the high interval of inter-item correlation ($0.5 \leq r_{xy} < 0.8$) were not found that. According to the high interval of inter-item correlation ($0.5 \leq r_{xy} < 0.8$) were found that Rudner's method may be significantly different with Lee's method at the significance level of .05 as shown in Table 6.

In consideration all cases of the sample sizes conditions were found that the classification indices estimated by Rudner's method non-significantly difference with Lee's method at the significance level of .05 as shown in Table 7.

Table 6. Comparisons of the Accuracies and Consistencies of the Indices Estimation Methods of Rudner and Lee in the Different Demonstration Situations

indices	**inter-item correlations**	n	U*	SE	Sig.
accuracy	$0.1 \leq r < 0.5$	100	28.0	7.826	0.710
		500	20.0	7.826	0.620
		1000	26.0	7.826	0.902
		5000	20.5	7.818	0.620
	$0.5 \leq r < 0.8$	100	7.0	7.826	**0.026***
		500	19.0	7.826	0.535
		1000	11.0	7.826	0.097
		5000	5.0	7.826	**0.011***
consistency	$0.1 \leq r < 0.5$	100	28.0	7.826	0.710
		500	20.0	7.826	0.620
		1000	26.0	7.826	0.902
		5000	22.0	7.826	0.805
	$0.5 \leq r < 0.8$	100	9.0	7.826	0.053
		500	19.0	7.826	0.535
		1000	11.0	7.826	0.097
		5000	5.0	7.826	**0.011***

Note: *Mann-Whitney U Test

Table 7. Comparisons of the Accuracies and Consistencies of the Indices Estimation Methods of Rudner and Lee with the Demonstrated Samples

indices	**Sample sizes (n)**	**U***	**SE**	**Sig.**
accuracy	100	77.5	21.761	0.352
	500	78.0	21.761	0.376
	1000	80.0	21.764	0.427
	2000	59.5	21.761	0.077
consistency	100	82.0	21.764	0.482
	500	78.0	21.764	0.376
	1000	80.0	21.764	0.408
	2000	61.0	21.764	0.094

Note: *Mann-Whitney U Test

Conclusions

The conclusions and discussions issues were as follows.

1. The Inter-Item Correlations and the Two Classification Indices Estimation Methods

By considering the accuracies and consistencies of the estimation methods with the empirical data in general and specific terms, it was found that Lee's method had higher accuracies and consistencies than Rudner's method. It can be stated that by controlling the influences of theta, the items are not related. This is consistent with the initial agreement of the index estimation about the standard error distribution. By analyzing the empirical data, it was found that the distribution was not normal distribution and inconsistent with Rudner's method. However, it was consistent with Lee's method (Wyse & Hao, 2012).

By considering the accuracies and consistencies of the estimation methods with the demonstration that had different correlations, both methods might have different items. With the correlation coefficient of $0.1 \leq r_{xy} < 0.5$, the two methods might be non-significantly different at the significance level of .05. With the correlation coefficient of $0.5 \leq r_{xy} < 0.8$, the two methods might be significantly different at the significance level of .05. Even though Rudner's method applies the maximum likelihood and Lee's method uses the integration, both methods use theta for the estimations (Lathrop & Cheng, 2013).Since the independences between the items affected the parameter estimation, the performances of the examinees affect the index estimation. Hence, these two methods may be significantly different.

2. The Sample Sizes and the Two Classification Indices Estimation Methods

By considering the different accuracies and consistencies of the indices estimations with the methods of Rudner and Lee as well as the different sample sizes regardless of the independencies of the items, it was found that both estimation methods had insignificant different accuracies and consistencies at the significance level .05. This was consistent with the initial investigation of the different sample sizes. The study of Wyse & Hao (2012) with the sample size of 2,000 units had similar estimation results with the studies having the sample sizes of 10,000 and 25,000 units.

Suggestions

1. Suggestions for Generalization

Since it was found that different samples regardless of the correlations between the items, the methods of Rudner and Lee had insignificant classification accuracies and consistencies at the significance level .05. Hence, any classification method can be used, even though this demonstration is not a normal distribution and consistent with the initial agreement for Rudner's method.

2. Suggestions for Further Studies

This simulation study had finding about the selection of the classification indices estimation methods in the case of the independencies between the items and different sample sizes. The obtained information are the classification accuracies and consistencies. Further studies may examine estimation biases and standard deviations of the classification indices estimations in order to obtain more and useful information for making decisions to use classification indices estimation methods in the future.

ACKNOWLEDGEMENTS

This research was financially sponsored by THE 90TH ANNIVERSARY OF CHULALONGKORN UNIVERSITY FUND (Ratchadaphiseksomphot Endowment Fund)

References

Guo, F. (2006). Expected classification accuracy using the latent distribution. Practical assessment. *Research & Evaluation, 11*(6), 1-9.

Houts, C. R., & Cai, L. (2013). *flexMIRT user's manual version* 2: *Flexible multidimensional item analysis and test scoring*. NC: Vector Psychometric Group.

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

Huynh, H. (1990). Computation and statistical inference for decision consistency indexs based on the Rasch model. *Journal of Educational Statistics*, *15*(4), 353-368.

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement, 37*(3), 226-241.

Lathrop, Q. N. (2015). cacIRT: *Classification Accuracy and Consistency under Item Response Theory. R package version 1.4.* http://CRAN.R-project.org/package=cacIRT

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31*(4), 255-274.

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*, 1-17.

Lee, W., Brennan, R. L., & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement, 33*(5), 374-390.

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.

Rudner, L. M. (2005). Expected classification accuracy. Practical Assessment, Research & Evaluation, 10(13), 1-4.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*(4), 265-276.

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

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement, 36*, 602-624.