mathematics

Article

The Utility of Receiver Operating Characteristic Curve in Educational Assessment: Performance Prediction

Hyunsuk Han

Department of Counseling Psychology, Kyungil University, Gyeongsan 38428, Korea; hh03318@kiu.kr

Abstract: When examinees are classified into groups based on scores from educational assessment, two indices are widely used to gauge the psychometric quality of the classifications: accuracy and consistency. The two indices take correct classifications into consideration while overlooking incorrect ones, where unbalanced class distribution threatens the validity of results from the accuracy and consistency indices. The single values produced from the two indices also fail to address the inconsistent accuracy of the classifier across different cut score locations. The current study proposed the concept of classification quality, which utilizes the receiver operating characteristics (ROC) graph to comprehensively evaluate the performance of classifiers. The ROC graph illustrates the tradeoff between benefits (true positive rate) and costs (false positive rate) in classification. In this article, a simulation study was conducted to demonstrate how to generate and interpret ROC graphs in educational assessment and the benefits of using ROC graphs to interpret classification quality. The results show that ROC graphs provide an efficient approach to (a) visualize the fluctuating performance of scoring classifiers, (b) address the unbalanced class distribution issue inherent in the accuracy and consistency indices, and (c) produce more accurate estimation of the classification results.

Keywords: assessment; classification quality; ROC graph; accuracy; consistency

MSC: 97D60

check for updates

Citation: Han, H. The Utility of Receiver Operating Characteristic Curve in Educational Assessment: Performance Prediction. *Mathematics* **2022**, *10*, 1493. https://doi.org/10.3390/math10091493

Academic Editors: David Pugalee, Gregorio Giménez Esteban and Javier Valbuena

Received: 2 March 2022
Accepted: 22 April 2022
Published: 30 April 2022

1. Introduction

The purpose of classification based on test results is to produce useful information regarding examinees. Since the use of cut scores to classify examinees is widely used, particularly in educational assessment, estimating the statistical performance of classifiers is a vital part of psychometric analysis. Classifications resulting from test cut scores can be dichotomous (e.g., pass or fail) or polytomous (e.g., far below benchmark, below benchmark, at benchmark, above benchmark) depending on the purpose and characteristics of the tests. Regardless, the essential purpose of the assessments is to produce interpretive meaning based upon an examinee's obtained score with reference scores (e.g., cut-off score) [1]. For widely used licensure tests (e.g., the Praxis tests, General Educational Development tests), pass/fail distinctions are often the final form of score reporting upon which stakes are based. High stakes decisions are also part of primary and secondary education. For example, the federal grant program entitled "Race to the Top" [2] emphasizes the accountability and instructional improvement of K-12 assessments, which typically result in classifying students by achievement and reflects the growing importance of improving the quality of classification and the interpretations based on classifications in education.

Classification using examinees' observed scores is an attempt to accurately categorize continuous quantities into several different groups, dependent on the cut-off scores. Classifying different groups of students in educational settings based on academic or social/behavioral assessments is critical in order to identify students in need of additional supports. For example, research suggests that students performing well below established achievement benchmarks, such as those used in the Dynamic Indicators of Early Literacy Skills (DIBELS) [3], are at risk for long-term negative outcomes, including grade retention and school drop-out [4]. Therefore, educators need accurate and reliable classification systems, based on assessment cut scores, to prevent negative outcomes and deliver necessary remediation to increase the likelihood of student success [5].

Cut-off scores have been successfully identified and used in educational assessments through various procedures [6–8]. Traditionally, the quality of the cut-off score has been evaluated by the use of accuracy and consistency of classification results with two indices: (a) classification accuracy (CA) and (b) classification consistency (CC) [9]. While reporting CA and CC is becoming a common practice, there has been limited research addressing some of the inherent limitations of the two indices. One of the most important threats to the appropriate use of CA and CC indices is unbalanced class distribution. In practice, the numbers of individuals in different classes can vary greatly, which causes problems when focusing exclusively on correct classes, as do the CC and CA indices [10].

Another limitation of the CA and CC indices is that they fail to show the varying performance of the classifier across the whole group. Let us consider the illustrated example in Table 1. In Case 1, a small number of examinees have actual failing status according to the cut score and yet the classifier placed all examinees into the passing group. CA would equal 80%, and this may be misleading because all examinees who should have been in the failing category were misclassified. In other words, the classifier does not perform well for examinees with lower abilities. In Case 2, the CA is again 80% and the distribution of actual classification groups is more balanced. However, the classifier does not perform well for examinees with higher abilities. Given that the goal of many educational assessments is to classify and identify a small group of students at risk, often to deliver necessary intervention, CA indices are concerning because the value can be the same even when the pattern of results are different, as was the circumstance in Case 1 and Case 2.

Table 1. An example showing the limitation of CA and CC indices.

Case 1	Predict:Pass	Predict:Fail	Case 2	Predict:Pass	Predict:Fail
Actual:Pass	80	0	Actual:Pass	30	20
Actual:Fail	20	0	Actual:Fail	0	50

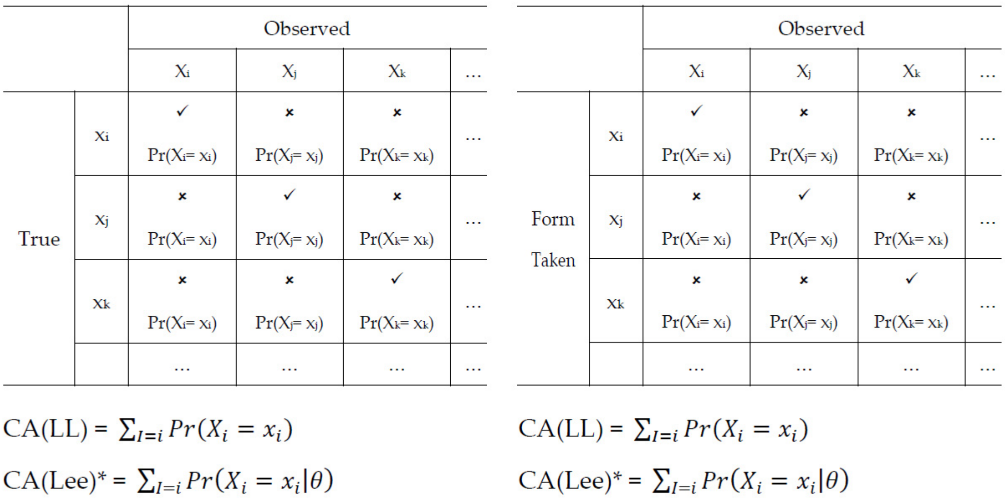
Since CC and CA are often poor metrics for measuring classification quality, other indices have been developed as alternatives. Specifically, the use of receiver operating characteristics (ROC) graphs is increasingly being used for estimating and visualizing the performance of classifiers [11,12]. Applying the ROC approach to educational assessment not only mitigates the limitations of CA and CC, but also provides more information about the performance of classifiers beyond CA and CC. However, few research have been conducted applying ROC despite the advantage of giving quality information regarding the classification. It is not as widely used to evaluate students' knowledge and skills [13] though the ROC has been widely used in various disciplines including medical or clinical settings for diagnostic purposes [14–17]. As such, the goal of this paper is to introduce the ROC graph as a means to overcome the limitations of CA and CC indices in educational assessment. Specifically, this study will demonstrate how the ROC approach is able to depict the tradeoff between benefits (i.e., true positive rate) and costs (i.e., false positive rate) in classification through simulation study. Below, the author will provide the theoretical background and use simulated data to show that the ROC functions well across various test conditions, including test length, sample size, and distribution of students' ability.

2. Theoretical Background

2.1. Classification Accuracy and Classification Consistency

Definition of Classification Accuracy and Classification Consistency

Classification accuracy (CA) is the rate at which the classifications based on observed cut scores agree with classifications based on true cut scores [9,18]. Classification consistency (CC) is the rate at which examinee classifications based on repeated independent and parallel test forms agree with each other [19]. There are two approaches that are commonly used to estimate CA and CC—the Livingston and Lewis approach [18], using the beta distribution, and the Lee approach [9], using the IRT framework—provided in Figure 1.



		Observed			
		Xi	Xj	Xk	…
True	xi	✓ Pr(Xi= xi)	× Pr(Xj= xj)	× Pr(Xk= xk)	…
	xj	× Pr(Xi= xi)	✓ Pr(Xj= xj)	× Pr(Xk= xk)	…
	xk	× Pr(Xi= xi)	× Pr(Xj= xj)	✓ Pr(Xk= xk)	…
		…	…	…	…

CA(LL) = $\sum_{I=i} Pr(X_i = x_i)$

CA(Lee)* = $\sum_{I=i} Pr(X_i = x_i|\theta)$

		Observed			
		Xi	Xj	Xk	…
Form Taken	xi	✓ Pr(Xi= xi)	× Pr(Xj= xj)	× Pr(Xk= xk)	…
	xj	× Pr(Xi= xi)	✓ Pr(Xj= xj)	× Pr(Xk= xk)	…
	xk	× Pr(Xi= xi)	× Pr(Xj= xj)	✓ Pr(Xk= xk)	…
		…	…	…	…

CA(LL) = $\sum_{I=i} Pr(X_i = x_i)$

CA(Lee)* = $\sum_{I=i} Pr(X_i = x_i|\theta)$

Figure 1. An example showing the CA and CC indices by Livingston and Lewis, and Lee, respectively. * *The probabilities of each response pattern in the Lee approach are computed as a given θ.*

The most widely used approach to estimate CA and CC is the Livingston and Lewis (LL) approach [9,20–24]. This approach has been used in a number of high stakes education assessment systems, including the California Standards Tests, the Florida Comprehensive Assessment Test, and the Washington Comprehensive Assessment Program.

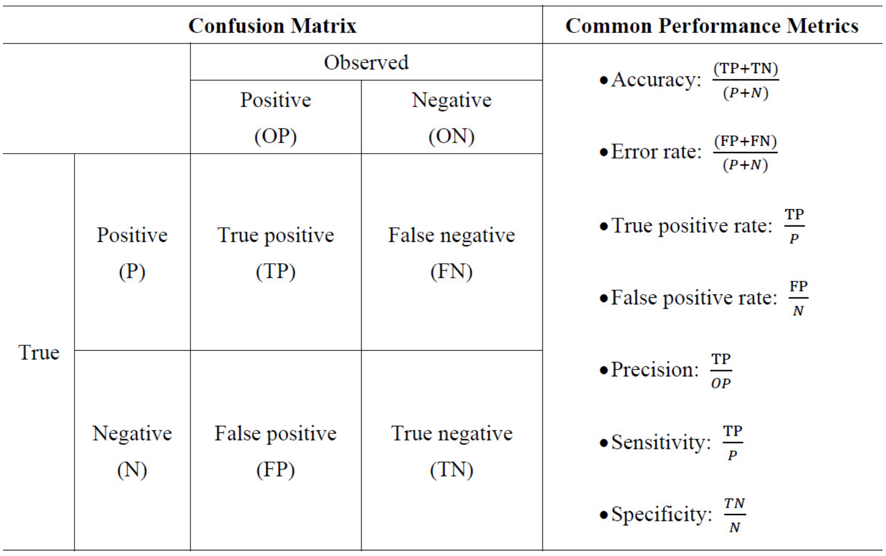
Another way of estimating CA and CC is the Lee's approach developed using the item response theory (IRT) framework [18,19,22,25,26]. Lee's approach has also been used in high stakes assessment systems, including the Connecticut Mastery Test. This approach first employs a compound multinomial distribution [27] to model the conditional summed-score distribution and aggregate the probabilities of scoring with a given category. Then, CA and CC are calculated using an n by n contingency table, similar to the LL approach.

2.2. Classification Matrix and ROC Graph

2.2.1. Classification Matrix

Classification matrix is constructed by using two classes, positive and negative, which can be regarded as "pass" and "fail" in real cases. The information on the true positive rate and the false positive rate can be computed using a 2-by-2 confusion matrix. Figure 2 shows a classification matrix and equations for several commonly used metrics that can be calculated from it. *True positive* signifies that a positive examinee is correctly classified as positive. *False negative* indicates that a positive examinee is misclassified as negative. *False positive* means that a negative examinee is misclassified as positive. Finally, *true negative* means that a negative examinee is correctly classified as negative.

There are more indices that can be estimated using the classification matrix, including negative predictive value, miss rate, false discovery rate and so on. This study focuses on the true positive rate $\frac{TP}{(TP+FN)}$ and the false positive rate $\frac{FP}{(FP+TN)}$ here as they are typically the values of interest. The true positive rate can be interpreted as the probability of positives that are correctly classified among all the positives, and the false positive rate can be interpreted as the probability of negatives that are incorrectly classified as positives among all the negatives.



Confusion Matrix

		Observed	
		Positive (OP)	Negative (ON)
True	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Common Performance Metrics

- Accuracy: $\frac{(\mathrm{TP}+\mathrm{TN})}{(P+N)}$
- Error rate: $\frac{(\mathrm{FP}+\mathrm{FN})}{(P+N)}$
- True positive rate: $\frac{\mathrm{TP}}{P}$
- False positive rate: $\frac{\mathrm{FP}}{N}$
- Precision: $\frac{\mathrm{TP}}{OP}$
- Sensitivity: $\frac{\mathrm{TP}}{P}$
- Specificity: $\frac{TN}{N}$

Figure 2. Confusion matrix and common performance metrics computed from it.

2.2.2. ROC Graph

ROC graphs are built with the true positive rate plotted on the *y*-axis and treated as a benefit, and false positive rate plotted on the *x*-axis and treated as a cost [28]. To estimate the classification quality in a test, it is necessary to plot costs and benefits in the ROC space. When examinees are classified in a test, only a single classification matrix can be generated based on the cut score, which corresponds to one single point in the ROC space. The single point represents the overall performance of a classifier.

In addition to the single point, this study also focuses on plotting a curve to reflect the performance of the classifier across different cut score locations. The ROC curve consists of the entire set of false positive rate and true positive rate pairs (i.e., cost–benefit) resulting from the continually changing cut scores over the range of test results, plotting the changing cut scores, and therefore it has been recognized as a global measure of a test's accuracy [29].

The cost–benefit approach uses the ROC graph to generalize information from the tradeoffs between false and true positives. Figure 3 shows the ROC graph where each point in the ROC space represents a classifier's performance.

A classifier at the lower left point O (0, 0) means that both cost and benefit are equal to 0. A classifier at the upper right point C (1, 1) means that both cost and benefit are equal to 1. A classifier at the upper left point B (0, 1) means that the cost is 0 and the benefit is 1, representing perfect classification. Intuitively, a classifier has better performance if its points in the ROC space are close to point B, where benefits are high, and costs are low. The diagonal line represents random performance because it means the classifier has a 50% chance of correctly classifying examinees into either positive or negative.

2.3. Use of the ROC to Estimate Classification Quality in Practice

The development of classification quality was inspired by the idea that both correctly classified examinees and those who are incorrectly classified should be considered to evaluate the performance of the classifier. Classification quality is the attribute of a classifier that portrays the relationship between the classifications based on observed cut scores and the classifications based on true cut scores. Accuracy and consistency indices have been computed for decades in the context of educational assessment, and the newly proposed classification quality serves as an alternative to alleviate some problems of the traditional indices.

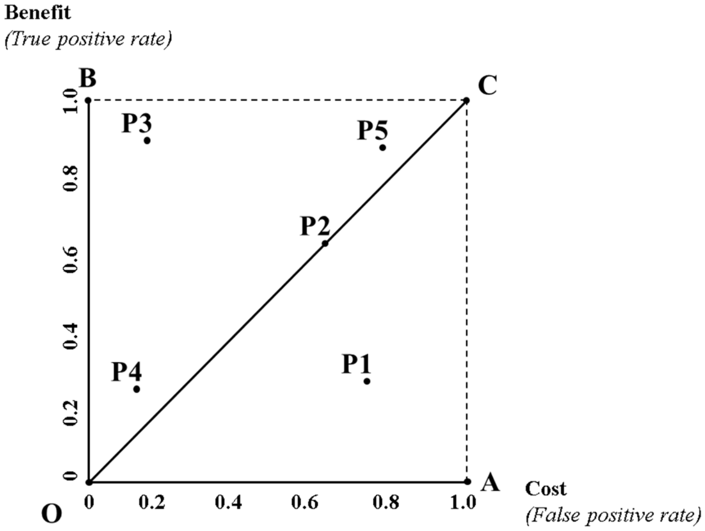


Figure 3. ROC graph and explanations of classifier positions.

One of the inherent problems of the accuracy and consistency indices is their unbalanced class distribution. In fact, any performance metric that uses values from multiple columns (e.g., "pass" and "fail") is inherently sensitive to class skews [30,31]. Even if the performance of the classifier does not change, the variation in class distribution changes the accuracy index because it is computed across multiple columns. ROC curves are different from other indices because they use a row-parallel computation using the classification. The true and false positive rates are ratios from two columns without crossing rows; therefore, the ROC curve is insensitive to class distributions. In practice, it is not unusual to see that less than 15% of respondents are classified into "fail" in many assessments based on a cut score [32], and this causes unbalanced class distribution. Another attractive feature of using ROC graphs is that they provide a tradeoff between costs and benefits across all cut scores in the sample. ROC graphs can uncover potential reasons behind the varied classification quality with the change in cut scores and can provide a visual display of the variation.

In this study, the area under the curve (*AUC*) is used to quantify cost–benefit. The *AUC* is the probability that a classifier will rank a randomly chosen positive instance higher than a negative one [33]. AUC is also equal to the value of the Wilcoxon-Mann–Whitney U test statistic. Basically, *AUC* is calculated based on the true positive rate and false positive rate. Formally, *AUC* given a cut score at *i* can be computed as Equation (1).

$$AUC = \frac{TPR * FPR + \frac{TPR(TPR+1)}{2} - R}{TPR * FPR} \tag{1}$$

where TPR denotes true positive rate $\frac{TP}{(TP+FN)}$, FPR denotes false positive rate $\frac{FP}{(FP+TN)}$, R_i denotes the sum of ranks. The AUC can be easier to calculate using the Gini coefficient [34] by $G_1 = 2AUC - 1$, given G_1 is computed as

$$G_1 = 1 - \sum_{k=1}^{n} (FPR_k - FPR_{k-1})(TPR_k + TPR_{k-1}) \tag{2}$$

Although *AUC* has a range between 0 to 1 in the *ROC* space, it is mentioned that a classifier should perform no worse than random guessing, which means that a classifier in real practice is expected to have an $AUC > 0.5$. In general, an AUC of 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding [35].

In summary, *ROC* graphs overcome the shortcomings of using accuracy indices and produce more detailed information on classification quality. In addition, multiple indices can be generated using classification matrices. Of those, *AUC* is a useful and informative indicator in cost–benefit analysis.

3. Method

A simulation study was conducted to investigate factors that might influence classification quality. Based on the literature reviews in psychometric simulation studies, factors were manipulated including sample size, test length, and distribution of ability [36–40]. In this study, the factors are specified as follows: test length (i.e., 20, 40, and 60 items), ability distributions (i.e., a normal distribution, a skewed distribution, and a mixed distribution), and sample size (i.e., 500, 1000, and 3000). Each condition was replicated 50 times.

In order to reflect practical settings, the true item parameters for the various simulated tests were taken from an operational, high-stakes test in the state of Florida. All items are binary scored, and the relationship between item parameters and the probability of correct response followed a 3-parameter logistic (3PL) model with difficulty, discrimination, and lower-asymptote parameters, $P_j(\theta) = c_j + (1 - c_j)/[1 + \exp(-1.7\alpha_j(\theta - \beta_j))]$ [41]. Specifically, 6th grade mathematic results are used to generate the simulated item parameters. The parameter estimation results are reported in an annual technical report (Florida Comprehensive Assessment Test, 2006), and item parameter estimate is summarized in Table 2.

Table 2. Operational item parameter and five-point summary and range.

Parameter	Min	25th Percentile	50th Percentile	75th Percentile	Max
discrimination	0.44	0.72	0.95	1.10	1.39
difficulty	−1.82	−0.73	−0.20	0.55	1.60
lower-asymptote	0.07	0.17	0.20	0.24	0.33

The range of the item discrimination estimates is from approximately 0.44 to 1.39. For difficulty parameters, the range is from −1.82 to 1.6. For lower-asymptote parameter, the range is 0.07 to 0.33. Using these parameter estimates, we generated the item parameters for the simulation. Specifically, the discrimination parameter was sampled from a uniform distribution with a range of (0.44, 1.39). The difficulty parameter was sampled from a uniform distribution with a range of (−1.82, 1.60). For the lower-asymptote parameter, the parameter was sampled from a uniform distribution with a range of (0.07, 0.33).

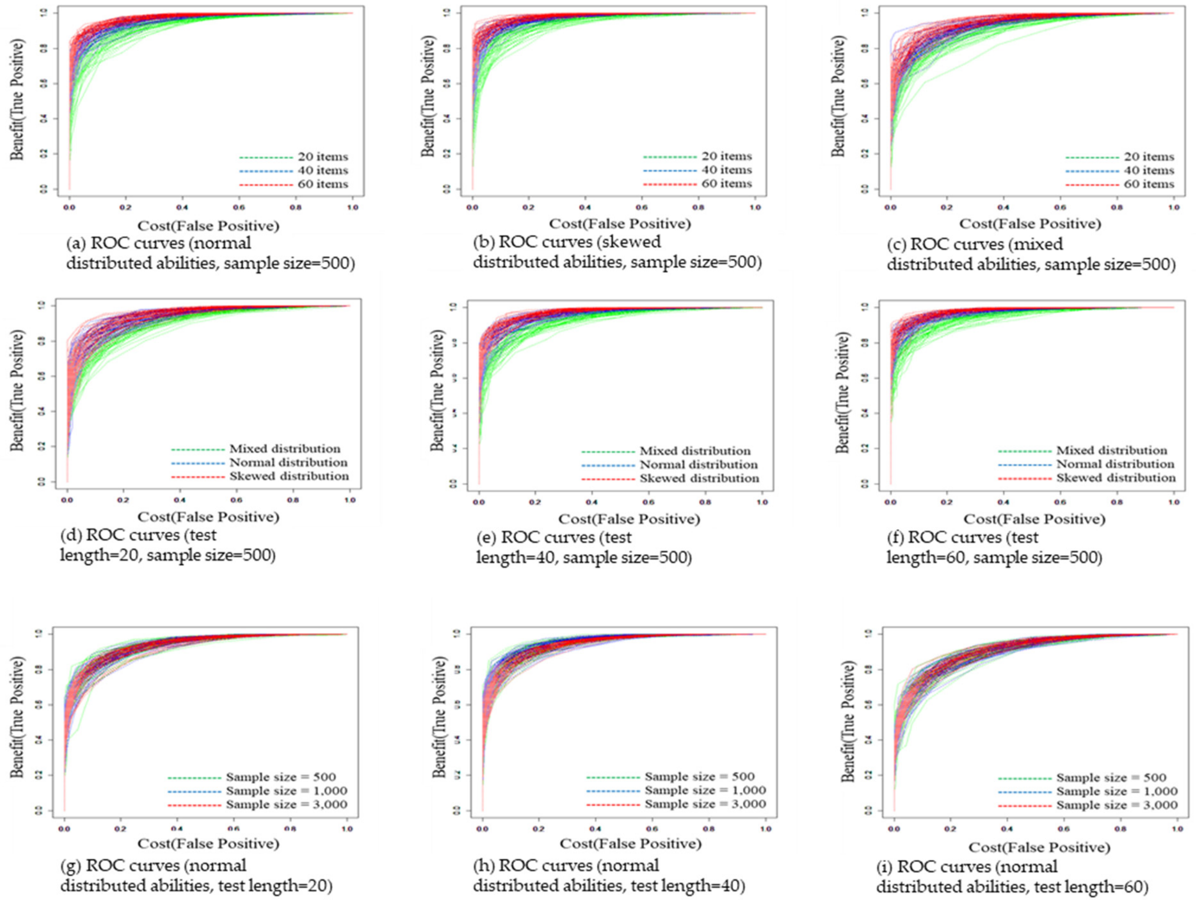
Abilities were drawn from three types of different ability distributions: (a) normal distribution, (b) Fleishman distribution, and (c) mixed distribution. Under the normal distribution, true ability followed a standard normal distribution. The data generation under the skewed condition followed Fleishman [37], where the ability distribution has mean 0, standard deviation 1, skewness 0.75, and kurtosis 0. In the mixed condition, ability came from two normal distributions $\mathcal{N}(-0.25, 0.61)$ and $\mathcal{N}$ (2.19, 1.05) with mixing proportion of 90% and 10%, respectively, which also has a mean 0 and standard deviation 1, as used in Woods [40].

To form ROC curves, results within each individual data set were aggregated across the 50 replications for each condition. Threshold averaging was used to aggregate across replications where both true positive rates and false positive rates were averaged at fixed intervals [42]. AUC were also calculated and reported for all ROC graphs.

Using generated item parameters and abilities, the response vectors were generated using the R package *cacIRT*. Using generated response vectors, we performed AUC calculation and ROC construction in the R package *ROCR*. R version 4.0.2 was used to perform the simulation [43].

4. Results

Figure 4a–i show how the ROC curves change with different test lengths (i.e., 20, 40, and 60 items) and sample size (i.e., 500, 1000, and 3000) when the ability distributions are normal, skewed, and mixed, respectively. Further, the figures show the performances of the different classifiers, and since 50 replications per each condition were performed, 50 graphs are depicted for each condition because each graph represents a single replication. Thus, graphs are depicted in each plot.



(a) ROC curves (normal distributed abilities, sample size=500)

(b) ROC curves (skewed distributed abilities, sample size=500)

(c) ROC curves (mixed distributed abilities, sample size=500)

(d) ROC curves (test length=20, sample size=500)

(e) ROC curves (test length=40, sample size=500)

(f) ROC curves (test length=60, sample size=500)

(g) ROC curves (normal distributed abilities, test length=20)

(h) ROC curves (normal distributed abilities, test length=40)

(i) ROC curves (normal distributed abilities, test length=60)

Figure 4. ROC curves change with different conditions.

Figure 4a–c show the trend of ROC curves with different test lengths under different ability distributions. The three graphs show the effects of test length by comparing the performance among the classifiers. Figure 4d–f show the trend of ROC curves with different ability distributions under different test lengths. The three graphs show the effects of the ability distribution by comparing the performance among the classifiers. Figure 4g–i show the trend of ROC curves with different sample sizes under different ability distributions. The three graphs show the effects of sample size by comparing performance among the classifiers.

Specifically, the results of the comparison of the performances of the classifiers based on AUC and cut-off score with sensitivity and specificity are reported in Tables 3 and 4 in the following section.

Results for average AUCs through replications and how the AUC changes with test length, sample size, and ability distribution are presented in Table 3. Table 3 shows that the use of ROC works outstanding for all conditions with AUC as Han, Pei, and Kamber's [35] suggestions (i.e., all AUCs > 0.9). AUC is the largest when the ability distribution is skewed, the second largest when examinees' ability distribution is normal, and the smallest when the ability distribution is mixed. This suggests that the classifier works best when the ability distribution is skewed. Results also shows that AUC increases with longer test length, as confirmed with ANOVA through 50 replications. However, the impact of sample size was indistinguishable (which was, again, confirmed with ANOVA procedures).

Although AUC indices give us the comprehensive performance of the classifier under each condition, the specific differences under each condition can be better seen in Table 4 in the following section.

Table 3. AUC indices by test length, sample size, and ability distribution.

Test Length	Sample Size	Normal	Skewed	Mixed
20	500	0.938 (0.012)	0.949 (0.011)	0.909 (0.019)
	1000	0.939 (0.012)	0.945 (0.011)	0.902 (0.016)
	3000	0.937 (0.011)	0.948 (0.010)	0.904 (0.014)
40	500	0.959 (0.010)	0.972 (0.009)	0.934 (0.015)
	1000	0.961 (0.010)	0.972 (0.007)	0.931 (0.012)
	3000	0.960 (0.005)	0.971 (0.006)	0.932 (0.009)
60	500	0.973 (0.005)	0.981 (0.005)	0.949 (0.010)
	1000	0.972 (0.005)	0.981 (0.003)	0.946 (0.007)
	3000	0.972 (0.004)	0.980 (0.003)	0.948 (0.007)

Table 4. Cut-off score with sensitivity and specificity by test length, sample size, and ability distribution.

Test Length	Sample Size	Normal			Skewed			Mixed		
		Cut-Off Score	Sensitivity	Specificity	Cut-Off Score	Sensitivity	Specificity	Cut-Off Score	Sensitivity	Specificity
20	500	0.582	0.861	0.854	0.589	0.875	0.859	0.586	0.831	0.801
	1000	0.588	0.867	0.842	0.590	0.888	0.859	0.584	0.830	0.820
	3000	0.589	0.868	0.847	0.591	0.883	0.862	0.588	0.839	0.810
40	500	0.554	0.906	0.887	0.556	0.916	0.902	0.552	0.872	0.841
	1000	0.554	0.900	0.889	0.558	0.916	0.900	0.554	0.864	0.839
	3000	0.553	0.897	0.883	0.558	0.916	0.900	0.556	0.864	0.837
60	500	0.542	0.924	0.904	0.542	0.935	0.923	0.543	0.881	0.860
	1000	0.542	0.921	0.907	0.543	0.930	0.921	0.543	0.881	0.865
	3000	0.542	0.917	0.901	0.542	0.929	0.918	0.544	0.886	0.861

The results in Table 4 provide the cut-off score, sensitivity, and specificity. One of the most distinctive trends is that when the number of items is increased, the cut-off score tends to decrease. Additionally, sensitivity and specificity are increased. This trend is applicable to all ability distributions. However, when the number of items is the same, the effect of sample size is indistinguishable. Nonetheless, sensitivity and specificity are the highest when the ability distribution is skewed, and lowest when the ability distribution is mixed.

5. Discussion and Conclusions

Assessment is the systematic process of implementing empirical data to measure knowledge, skills, attitudes, and beliefs. This study introduces a cost–benefit approach that overcomes the problems of using CC and CA methods, as well as provides practitioners with more information about classification quality. The author demonstrated two ways of using the cost–benefit approach to estimate classification quality: (a) plot the ROC point generated by the classification in the ROC space to show the classification performance, and (b) build the ROC curve to show the classification performance across all cut score locations. It is worth mentioning that neither way requires additional statistical analysis of the original dataset, and practitioners with the classification results can easily carry it out. A simulation study was conducted on classification quality. The results show that (a) the use of ROC methodology works well for classification, and (b) longer tests usually produce higher classification quality.

It should be noted that true classification distinctions should be given to applying ROC for evaluating classification quality in practice. For example, one can identify true classification results in marketing settings (e.g., whether or not a customer purchases a product) that are directly observable. This can also happen in medical or clinical settings where one can confirm that a patient has a disease or not through surgery or a pre-existing valid diagnostic instrument. Given a priori decision on classification, ROC can be used in education for evaluating the performance of classifiers under various conditions or among several different methods, as well as determining the optimal cut-point of a particular test. However, unlike other fields, such as marketing, it is challenging to know for certain whether or not examinees truly possess the trait that is intended to be measured in educational or psychological settings. Since many of these traits (e.g., language ability) are unobservable, it is not always possible to know true classification results a priori in practice unless the data come from simulated datasets. This is the fundamental problem of applying ROC in situations of educational assessment or licensure tests.

Notwithstanding the limitations of applying ROC in practice, there are additional research questions to address the usefulness of using ROC in education testing contexts. First, true classification specification might be achieved through cognitive diagnostic modeling (CDM), which leads to the potential use of ROC in determining the optimal cutoff scores of a particular assessment. As long as we can specify reliable confirmed diagnostic results, ROC can be used for evaluating the performance of the assessment as well as determining cutoff scores depending on the purpose of the assessment. Second, in the field of vocational education, it is quite feasible to determine whether a person has a particular skill (i.e., psycho-motor skills) or not regardless of those licensure tests. Based on judgment from true classification results, ROC can be used for evaluating the performance of the assessment as well as determining cutoff scores. In sum, when we want to apply ROC dealing with a latent trait for both scenarios, the core factor is to determine how one can specify whether people truly possess the trait. Future research should consider these aspects and conduct research on the performance of ROC in relation to CDM or psycho-motor skills and how and to what extent we can assure the accuracy of those results.

There are other directions that we would like to address in future research as well. Since there are multiple ways to generate the ROC curve, future research should consider other ways to fit a curve that better estimates classification quality. Another is that binary classification is used to illustrate the ROC curve in this study. Multi-class AUC and ROC curves have been developed in the area of machine learning. It is worth expanding current research to multi-class instances because many educational assessments use multi-classification approaches.

As noted, classifying different groups of students in educational settings based on academic or social/behavioral assessments is critical in order to identify students in need of additional supports. It is believed that the cost–benefit approach described here can help educational researchers and practitioners increase the accuracy of their classification approaches when assessing student performance. Classification of students and teachers is important for making educational and professional decisions, particularly with regard to identification of those in need of early intervention to increase the likelihood of positive future outcomes. By using a ROC approach, educators can ensure that they are correctly identifying those in need and increase broader confidence in the accuracy of their assessments.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Breyer, F.J.; Lewis, C. Pass-Fail Reliability for Tests with Cut Scores: A Simplified Method. *ETS Res. Rep. Ser.* **1994**, *1994*, 253–264. [CrossRef]
2. U.S. Department of Education. *Race to the Top Program Executive Summary*; U.S. Department of Education: Washington, DC, USA, 2009.
3. Good, R.H.; Kaminski, R.A. *Dynamic Indicators of Basic Early Literacy Skills*, 6th ed.; Institute for the Development of Educational Achievement: Eugene, OR, USA, 2002; Available online: http://dibels.uoregon.edu (accessed on 2 March 2022).
4. Duncan, G.J.; Dowsett, C.J.; Claessens, A.; Magnuson, K.; Huston, A.C.; Klebanov, P.; Pagani, L.S.; Feinstein, L.; Engel, M.; Brooks-Gunn, J.; et al. School readiness and later achievement. *Dev. Psychol.* **2007**, *43*, 1428–1446. [CrossRef] [PubMed]
5. Partanen, M.; Siegel, L.S. Long-term outcome of the early identification and intervention of reading disabilities. *Read. Writ.* **2013**, *27*, 665–684. [CrossRef]
6. Angoff, W.H. Scales, Norms and Equivalent Scores. In *Educational Measurement*, 2nd ed.; Thorndike, R.L., Ed.; American Council on Education: Washington, DC, USA; pp. 508–600.
7. Ferrara, S.; Perie, M.; Johnson, E. *Setting Performance Standards: The Item Descriptor (ID) Matching Procedure*; American Educational Research Association: New Orleans, LA, USA, 2002.
8. Lewis, D.M.; Green, D.R.; Mitzel, H.C.; Baum, K.; Patz, R.J. *The Bookmark Standard Setting Procedure: Methodology and Recent Implementations*; Council for Measurement in Education: San Diego, CA, USA, 1998.
9. Livingston, S.A.; Lewis, C. Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *J. Educ. Meas.* **1995**, *32*, 179–197. [CrossRef]
10. Lewis, D.; Gale, W. A sequential algorithm for training text classifiers. In Proceedings of the SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, Dublin, Ireland, 3–6 July 1994; Springer: Heidelberg, Germany; Dublin, Ireland, 1994.
11. Swets, J.A.; Dawes, R.M.; Monahan, J. Better Decisions through Science. *Sci. Am.* **2000**, *283*, 82–87. [CrossRef] [PubMed]
12. Zhang, J.; Mueller, S.T. A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika* **2005**, *70*, 203–212. [CrossRef]
13. Muñoz-Repiso, A.G.V.; Tejedor, F.J.T. The incorporation of ICT in higher education. The contribution of ROC curves in the graphic visualization of differences in the analysis of the variables. *Br. J. Educ. Technol.* **2012**, *43*, 901–919. [CrossRef]
14. Greiner, M.; Pfeiffer, D.; Smith, R. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev. Vet. Med.* **2000**, *45*, 23–41. [CrossRef]
15. Krach, S.K.; McCreery, M.P.; Wang, Y.; Mohammadiamin, H.; Cirks, C.K. Diagnostic Utility of the Social Skills Improvement System Performance Screening Guide. *J. Psychoeduc. Assess.* **2017**, *35*, 391–409. [CrossRef]
16. Quirk, M.; Dowdy, E.; Dever, B.; Carnazzo, K.; Bolton, C. Universal School Readiness Screening at Kindergarten Entry. *J. Psychoeduc. Assess.* **2018**, *36*, 188–194. [CrossRef]
17. Walter, S.D. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat. Med.* **2002**, *21*, 1237–1256. [CrossRef] [PubMed]
18. Lee, W. Classification consistency and accuracy for complex assessments using item response theory. *J. Educ. Meas.* **2010**, *47*, 1–17. [CrossRef]
19. Lee, W.-C.; Hanson, B.A.; Brennan, R.L. Estimating Consistency and Accuracy Indices for Multiple Classifications. *Appl. Psychol. Meas.* **2002**, *26*, 412–432. [CrossRef]
20. Berk, R.A. Selecting the index of reliability. In *A Guide to Criterion Referenced Test Construction*; Berk, R.A., Ed.; Johns Hopkins University Press: Baltimore, MD, USA, 1984.
21. Hanson, B.A.; Brennan, R.L. An Investigation of Classification Consistency Indexes Estimated Under Alternative Strong True Score Models. *J. Educ. Meas.* **1990**, *27*, 345–359. [CrossRef]
22. Huynh, H. Statistical consideration of mastery scores. *Psychometrika* **1976**, *41*, 65–78. [CrossRef]
23. Lee, W.; Brennan, R.L.; Wan, L. Classification consistency and accuracy for complex assessments under the compound multinomial model. *Appl. Psychol. Meas.* **2009**, *33*, 374–390. [CrossRef]
24. Subkoviak, M.J. Estimating Reliability from A Single Administration of A Criterion-Referenced Test. *J. Educ. Meas.* **1976**, *13*, 265–276. [CrossRef]
25. Schulz, E.M.; Kolen, M.J.; Nicewander, W.A. A Rationale for Defining Achievement Levels Using IRT-Estimated Domain Scores. *Appl. Psychol. Meas.* **1999**, *23*, 347–362. [CrossRef]
26. Wang, T.; Kolen, M.J.; Harris, D.J. Psychometric Properties of Scale Scores and Performance Levels for Performance Assessments Using Polytomous IRT. *J. Educ. Meas.* **2000**, *37*, 141–162. [CrossRef]
27. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking: Methods and Practices*, 2nd ed.; Springer: New York, NY, USA, 2004.
28. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E.R. Small-sample precision of ROC-related estimates. *Bioinformatics* **2010**, *26*, 822–830. [CrossRef]
29. McClish, D.K. Analyzing a Portion of the ROC Curve. *Med. Decis. Mak.* **1989**, *9*, 190–195. [CrossRef] [PubMed]
30. Fawcett, T. Using rule sets to maximize ROC performance. In Proceedings of the 2001 IEEE international conference on data mining, San Jose, CA, USA, 29 November–2 December 2001; pp. 131–138.
31. Fawcett, T. An Introduction to ROC analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861–874. [CrossRef]

32. Gitomer, D.H.; Qi, Y. *Recent Trends in Mean Scores and Characteristics of Test-Takers on Praxis II Licensure Tests*; U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service: Washington, DC, USA, 2010. Available online: https://www2.ed.gov/rschstat/eval/teaching/praxis-ii/report.pdf (accessed on 2 March 2022).
33. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
34. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Thomson Wadsworth: Belmont, CA, USA, 1984.
35. Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2011.
36. De Ayala, R. The Nominal Response Model in Computerized Adaptive Testing. *Appl. Psychol. Meas.* **1992**, *16*, 327–343. [CrossRef]
37. Fleishman, A.I. A method for simulating non-normal distributions. *Psychometrika* **1978**, *43*, 521–532. [CrossRef]
38. Hulin, C.L.; Lissak, R.I.; Drasgow, F. Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study. *Appl. Psychol. Meas.* **1982**, *6*, 249–260. [CrossRef]
39. Ranger, J.; Kuhn, J.-T. Assessing Fit of Item Response Models Using the Information Matrix Test. *J. Educ. Meas.* **2012**, *49*, 247–268. [CrossRef]
40. Woods, C.M. Empirical Histograms in Item Response Theory with Ordinal Data. *Educ. Psychol. Meas.* **2007**, *67*, 73–87. [CrossRef]
41. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*; Lord, F.M., Novick, M.R., Eds.; Addison–Wesley: Reading, MA, USA, 1968.
42. Provost, F.; Domingos, P. Well-trained PETs: Improving probability estimation trees. In *CeDER Working Paper #IS-00-04*; Stern School of Business, New York University: New York, NY, USA, 2001.
43. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020; Available online: http://www.R-project.org/ (accessed on 2 March 2022).