2021, Vol. 21, No. 3
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba

The Influence of Item Discrimination on Misclassification of Test Takers

Influencia de la discriminación de los ítems en la clasificación incorrecta de los examinados

R. Emmanuel Trujano * [1]

1 - Dirección de Investigación, Calidad Técnica e Innovación Académica, Centro Nacional de Evaluación para la Educación Superior, Mexico City, Mexico.

Recibido: 30/04/2021 **Revisado:** 05/08/2021 **Aceptado:** 08/08/2021

Introduction
Method
Results
Discussion
References

Abstract

It has been suggested that low discriminating items can be included in a test with a criterion-referenced score interpretation as long as they measure a highly relevant content. However, low item discrimination increases the standard error of measurement, which might increase the expected proportion of misclassified test takers. In order to test it, responses from 2000 test takers to 100 items were simulated, varying item discrimination values and number and location of cut scores, and classification inaccuracy was estimated. Results show that the expected proportion of misclassified test takers increased as item discrimination decreased, and as the cut scores were closer to the mean of the distribution of test takers. Therefore, a test should include as few items with low discrimination values as possible —or even none— in order to reduce the expected proportion of test takers classified into a wrong performance level.

Keywords: *decision accuracy, item discrimination, item response theory, Rudner algorithm, information function*

Resumen

Se ha sugerido que en un examen con interpretación de puntajes basada en criterios se pueden incluir ítems con baja discriminación siempre que midan un contenido muy relevante. Sin embargo, los ítems con baja discriminación aumentan el error estándar de medición, lo que podría aumentar la proporción esperada de examinados mal clasificados. Para probarlo, se simularon las respuestas de 2000 examinados a 100 ítems, variando la discriminación de los ítems, el número y ubicación de los puntos de corte, y se estimó la imprecisión de la clasificación. Los resultados muestran que la proporción esperada de examinados mal clasificados aumentó a medida que disminuyó la discriminación de los ítems y que los puntos de corte se acercaron a la media de la distribución de los examinados. Por lo tanto, un examen debería incluir la menor cantidad posible de ítems con baja discriminación —o incluso ninguno— para reducir la proporción esperada de examinados clasificados en un nivel de desempeño incorrecto.

Palabras clave: *precisión de la decisión, discriminación de los ítems, teoría de respuesta al ítem, algoritmo de Rudner, función de información*

***Correspondence to:** R. Emmanuel Trujano, Dirección de Investigación, Calidad Técnica e Innovación Académica, Centro Nacional de Evaluación para la Educación Superior. Avenida Camino al Desierto de los Leones (Altavista) 37, San Ángel, Álvaro Obregón, C.P. 01000, Mexico City, Mexico. Tel.: (+52) 5528205622. E-mail: rmanute@gmail.com

How to cite: Trujano, R. E. (2021). The Influence of Item Discrimination on Misclassification of Test Takers. *Revista Evaluar*, *21*(3), 15-34. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar

Participaron en la edición de este artículo: Rita Hoyos, Andrea Suárez, Eugenia Barrionuevo, Alicia Molinari, Mónica Serppe, Stefano Macri, Florencia Ruiz, Benjamín Casanova, Ricardo Hernández.

Introduction

When a test taker is assessed using a test with a criterion-referenced score interpretation, its performance is referenced to a previously well-defined set of knowledge, skills, or abilities, congruent with the purpose of the test (Popham, 2014; Richaud de Minzi, 2008). If a cut score is set for this test, guidelines for test assembly within the framework of item response theory (IRT) are available, such as maximizing the test information function (TIF) around the cut score value (Lord, 1980), which is accomplished by selecting items with difficulty parameter estimates close to the cut score value and an item discrimination as high as possible (Luecht, 2016).

Another guideline sometimes suggested is to select items that measure contents judged as highly relevant by subject-matter experts (SMEs), even though all test takers —or none— answer them correctly and, therefore, their discrimination parameter estimates are low (Burton, 2001; Clifford, 2016; Frisbie, 2005; Haladyna, 2016; Popham & Husek, 1969). However, this suggestion should be carefully considered because lower item discrimination decreases TIF; since this is associated with an increase in the standard error of measurement and a subsequent increase in the test takers' expected classification inaccuracy (Cheng, Liu, & Behrens, 2015), the inclusion of items with low discrimination estimates may increase the expected proportion of test takers classified into a wrong performance level. Previous research seems to suggest this is the case (Lathrop & Cheng, 2013; Luecht, 2016; Xing & Hambleton, 2004), so the purpose of this research is to further test if item discrimination influences the expected proportion of misclassified test takers.

A simulation study was conducted in which item discrimination was manipulated and the expected proportion of misclassified test takers was recorded. Dichotomously scored test items were simulated because the multiple-choice item is the most used item type among many testing programs (Haladyna, Rodriguez, & Stevens, 2019), and responses are usually scored as correct or incorrect (Haladyna, 2016). In addition, the number of cut scores and their location relative to the test takers' ability distribution was also manipulated since previous research has shown that these factors influence the classification inaccuracy (Ercikan & Julian, 2002; Lathrop & Cheng, 2013; Lee, 2010; Martineau, 2007; Wyse & Hao, 2012). Finally, responses were simulated using either the one-parameter logistic (1PL) or the two-parameter logistic (2PL) IRT model because TIF values obtained with 1PL model are more constrained due to the discrimination value shared by all items whereas TIF values obtained with 2PL model are less constrained due to the variability of discrimination values across items (see Luecht, 2016). Therefore, the more constrained TIF values from 1PL should derive in more classification inaccuracies for 1PL than for 2PL.

Method

This section describes the simulated conditions and the steps I followed to conduct the simulations.

Test Takers Ability Distribution

Samples of 2000 test takers were drawn from a standard normal distribution with a mean of 0 and a standard deviation of 1, which is the a priori distribution used by programs such as BILOG-MG (see Luecht, 2016), IRTPRO (Paek

& Hang, 2013) or R package ltm (Rizopoulos, 2018). These parameters were fixed across all conditions.

Number and Location of Cut Scores

Simulations were conducted with either one or two cut scores. When one cut score was simulated, it could take one of the following values: -1, 0, and 1. Notice that the value of 0 overlaps with the mean of the test takers' ability distribution.

When two cut scores were simulated, the first was always fixed at -1.5 and the second could take one of the following values: 0, 1.5, and 3. Once again, the value of 0 overlaps with the mean of the test takers' ability distribution.

Item Parameters

Responses to 100 dichotomously scored test items were simulated using either the 1PL or the 2PL IRT model.

One Cut Score. When one cut score was simulated, 100 item difficulty values *b* were drawn from a normal distribution with a standard deviation of 1 and a mean equal to each of the cut score values (-1, 0, and 1). In order to reduce variability in the results associated to variability in item difficulty across conditions, the same item difficulties were used to simulate responses with the 1PL and 2PL models and estimate classification inaccuracy.

For the 1PL model, seven item discrimination values *a* were used in the simulations: .25, .5, .75, 1, 1.5, 2, and 2.5. For the 2PL model, 100 discrimination values were drawn from a lognormal distribution with a standard deviation of 1 and one out of seven means: -2, -1, -.5, 0, 1, 2, and 4, subject to the constraint that $0 \leq a \leq 3$. Discrimination values were selected following the classification suggested by Baker and Kim (2017), and DeMars (2010) for their interpretation.

It is important to point out that, once simulated, item discrimination values were paired manually with item difficulty values in such a way so as to maximize test information around the cut score. Specifically, item difficulties were sorted from lowest to highest, whereas item discriminations were sorted to pair the highest values with the difficulties closest to the cut score values. (This was done manually in order to be certain about the location of the maximum values of TIF, since there was a technical problem trying to accomplish it with code).

Table 1 shows descriptive statistics of the simulated difficulty and discrimination values.

Two Cut Scores. When two cut scores were simulated, 50 item difficulties per each cut score value were drawn from a normal distribution with a mean equal to the cut score values, a standard deviation of .5, and a range equal to the mean ± .75. Specifically, for the first cut score (-1.5), *b* was sampled from N (μ = -1.5, σ = .5, *min* = -2.25, *max* = -.75), and when the second cut score was 0, 1.5, or 3, *b* was sampled in the following fashion:

- cut score = 0, *b* was sampled from N (μ = 0, σ = .5, *min* = -.75, *max* = .75).

- cut score = 1.5, *b* was sampled from N (μ = 1.5, σ = .5, *min* = .75, *max* = 2.25).

- cut score = 3, *b* was sampled from N (μ = 3, σ = .5, *min* = 2.25, *max* = 3.75).

Table 1
Descriptive statistics of item parameters for simulations with one cut score.

Mean of sampled distribution				Percentiles (Median)					
	M	SD	Min	P10	P25	P50	P75	P90	Max
Difficulty *b*									
-1.0	-0.993	0.894	-3.437	-2.068	-1.573	-0.939	-0.471	0.219	1.420
0.0	0.025	0.992	-2.202	-1.176	-0.645	0.006	0.650	1.175	3.024
1.0	0.999	1.084	-1.885	-0.228	0.361	1.019	1.605	2.552	3.378
Discrimination *a*									
-2.0	0.264	0.335	0.017	0.051	0.075	0.154	0.307	0.620	2.460
-1.0	0.563	0.564	0.011	0.116	0.228	0.401	0.664	1.161	2.990
-0.5	0.798	0.639	0.057	0.158	0.281	0.641	1.166	1.722	2.529
0.0	1.101	0.676	0.095	0.305	0.552	1.008	1.573	2.140	2.832
1.0	1.501	0.742	0.126	0.600	0.861	1.507	2.177	2.531	2.985
2.0	1.930	0.771	0.224	0.877	1.310	2.024	2.607	2.837	2.996
4.0	2.308	0.574	0.567	1.444	2.001	2.482	2.728	2.905	2.997

These item difficulties were used to simulate responses with the 1PL and 2PL models and estimate classification inaccuracy in order to reduce variability in the results associated with variability in item difficulty across conditions.

For the 1PL model, the same seven item discrimination values a were used in the simulations: .25, .5, .75, 1, 1.5, 2, and 2.5. For the 2PL model, 50 discrimination values per cut score were drawn again from a lognormal distribution with a standard deviation of 1 and one out of seven means: -2, -1, -.5, 0, 1, 2, and 4, subject to the constraint that $0 \leq a \leq 3$. Once again, the highest discrimination values were manually paired with the item difficulties closest to the cut score values in order to maximize test information around the cut scores: for each cut score, item difficulties were sorted from lowest to highest, whereas item discriminations were sorted to pair the highest values with the difficulties closest to the cut score values. (Again, this was done manually in order to be certain about the location of the maximum values of TIF, since there was a technical problem trying to accomplish it with code).

In order to simulate responses to 100 dichotomously scored test items, the 50 item parameters centred at cut score = -1.5 were combined with the 50 item parameters centred at each of the remaining cut scores. Table 2 shows descriptive statistics of each combination of simulated difficulty and discrimination values.

In summary, a total of 2 (one or two cut scores) × 3 (cut score values) × 2 (IRT models) × 7 (item discrimination values) conditions were simulated. Within each condition, the expected proportion of misclassified test takers was calculated with the Rudner algorithm (Rudner, 2001, 2005), which assumes that an individual's estimated ability $\hat{\theta}$ follows a normal distribution with mean θ and standard error $SE_{(\theta)} = 1/\sqrt{I(\theta)}$ (that is, the inverse of the square root of TIF). If an individual's estimated ability is below the cut score value, the probability of misclassification is the area under the normal distribution which is above the cut score. Conversely, if an individual's estimated ability is above the cut score value, the probability of misclassification is the area under the normal distribution which is below the cut score. The expected proportion of misclassified

Table 2
Descriptive statistics of item parameters for simulations with two cut scores.

Mean of sampled distribution				Percentiles (Median)					
	M	SD	Min	P10	P25	P50	P75	P90	Max
Difficulty *b* [a]									
$\mu_{cs2} = 0.0$	-0.735	0.839	-2.235	-1.898	-1.483	-0.729	-0.045	0.350	0.606
$\mu_{cs2} = 1.5$	0.024	1.555	-2.235	-1.898	-1.483	0.030	1.509	1.805	2.230
$\mu_{cs2} = 3.0$	0.753	2.272	-2.235	-1.898	-1.483	0.800	2.991	3.337	3.629
Discrimination *a*									
-2.0	0.245	0.280	0.005	0.048	0.077	0.154	0.308	0.570	1.874
-1.0	0.592	0.542	0.023	0.137	0.237	0.410	0.745	1.267	2.772
-0.5	0.713	0.493	0.041	0.159	0.340	0.633	1.022	1.467	2.302
0.0	1.136	0.688	0.105	0.334	0.599	1.048	1.573	2.182	2.922
1.0	1.557	0.736	0.229	0.598	0.961	1.521	2.111	2.613	2.876
2.0	1.974	0.634	0.526	1.141	1.436	2.027	2.519	2.792	2.982
4.0	2.296	0.631	0.410	1.390	1.961	2.494	2.821	2.932	2.997

Note. [a]The table shows descriptive statistics of 100 item difficulties sampled from two normal distributions: 50 from a distribution with a mean of $\mu_{cs1} = -1.5$ (the first cut score), and 50 from a distribution with a mean equal to each value of the second cut score (μ_{cs2}).

test takers is the average across individuals of the probabilities of misclassification.

Data Generation Steps

Within each condition, data were generated as follows:

1. Set the number of cut scores.
2. Set the cut score values.
3. Set the item parameter values.
4. Draw a sample of 2000 test takers from a standard normal distribution.
5. Simulate 100 responses to dichotomously scored test items with either the 1PL or the 2PL IRT model.
6. Estimate the test takers' maximum likelihood ability and their standard error of measurement according to their simulated responses and the item parameter values.
7. For each test taker, estimate the probability of misclassification, that is, the probability of being classified into performance level B when its estimated ability level falls into performance level A (p(B|A)).
8. Estimate the overall expected classification inaccuracy by averaging all the individual probabilities of misclassifications.
9. Estimate the expected proportion of false positives and false negatives when a single cut score was simulated, or the expected proportion of each misclassification when two cut scores were simulated, by averaging the corresponding individual probabilities of misclassifications.

10. Repeat steps 4 to 9 1000 times.

All the simulations were conducted in the R statistical software (R Core Team, 2020): Item parameters were simulated with package Runuran (Leydold & Hörmann, 2021); responses to items were simulated and test takers' abilities were estimated with package irtoys (Partchev, Maris, & Hattori, 2017); data were plotted with package ggplot2 (Wickham, 2016; Wickham et al., 2021); and the expected proportions of misclassified test takers were estimated with code adapted from package cacIRT (Lathrop, 2014, 2015). Appendix 1 shows the code used to conduct simulations with one cut score, whereas Appendix 2 shows the code for simulations with two cut scores.

Results

Figure 1 shows boxplots of the expected proportion of misclassified test takers as a function of item discrimination for simulations with one cut score using 1PL IRT model. The general trend across all panels is that the expected misclassification values decrease as item discrimina-

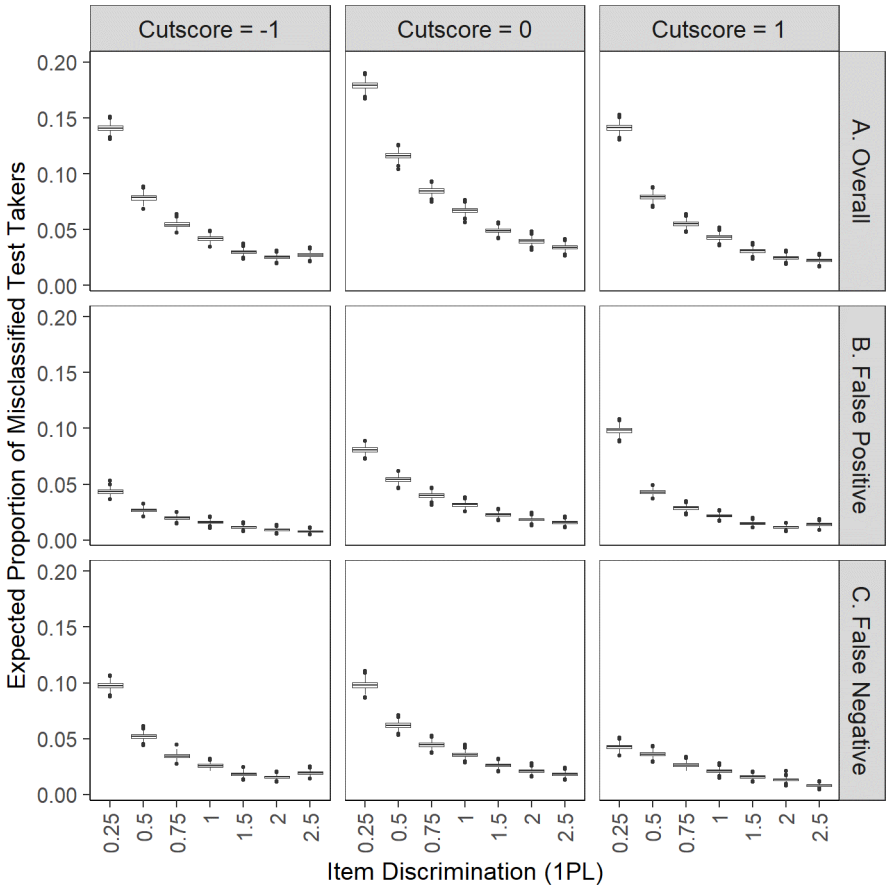

Figure 1
Expected proportion of misclassified test takers as a function of item discrimination for simulations with one cut score using 1PL IRT model.

Note. Data points represent outliers within each condition.

tion increases.

The top panels show that the overall expected classification inaccuracies are higher when the cut score is placed in the mean of the test takers' ability distribution, irrespective of item discrimination values, and lower otherwise. Compare, for example, the boxplots of overall misclassifications in the top panels when item discrimination equals .25.

With one cut score, misclassifications are of two types: a false positive (being wrongly classified into the higher performance level when estimated ability score belongs to the lower one) and a false negative (being wrongly classified into the lower performance level when estimated ability score belongs to the higher one). The middle and bottom panels of Figure 1 show expected misclassification values separated into false positives and negatives, respectively. Within each item discrimination value, false positives are comparable when cut score equals 0 and 1, and lower when cut score equals -1; as an example, compare the boxplots of false positives in the middle panels when item discrimination equals .25. In contrast, false negatives are comparable for each item discrimination value when cut score equals -1 and 0, and lower when cut score equals 1; for example, compare the boxplots of false negatives in the bottom panels when item discrimination equals .25.

Notice that all the distributions are narrow, and even the data points corresponding to outliers at each boxplot are not far away between each other. This implies that the expected misclassification values are consistent within each condition.

Figure 2 shows boxplots of the expected proportion of misclassified test takers as a function of item discrimination for simulations with one cut score using 2PL IRT model. The same trends as those using 1PL model can be observed: the expected misclassification values decrease as the median of item discrimination increases across all panels, the overall expected classification inaccuracies when the cut score is placed in the mean of the test takers' ability distribution are higher than when it is placed farther (irrespective of the median of item discrimination values), false positives are lower when cut score equals -1 than when it equals 0 and 1, false negatives are lower when cut score equals 1 than when it equals -1 and 0, and all the distributions are narrow (implying that expected misclassification values are consistent within each condition).

When Figures 1 and 2 are compared, the proportions of misclassified test takers are lower for the 2PL than for the 1PL model. As an instance, compare the top panels of both figures: the maximum overall expected misclassifications reach a median of .15 for 2PL when the cut score equals 0, lower than the median of approximately .18 for 1PL.

Figure 3 shows boxplots of the expected proportion of misclassified test takers as a function of item discrimination for simulations with two cut scores using 1PL IRT model. Notice the different scales on the *y*-axis. Consistent with the results of simulations with one cut score, the expected misclassification values decrease as item discrimination increases.

The top panels show that the overall expected classification inaccuracies decrease as the value of the second cut score shifts away from the mean of the test takers' ability distribution, but lower item discrimination is still associated to higher expected misclassification of test takers.

With two cut scores and three performance levels, there are two misclassification types per each performance level: a false performance level 2 and a false performance level 3 when estimated ability score belongs to performance level 1, a false performance level 1 and a false performance

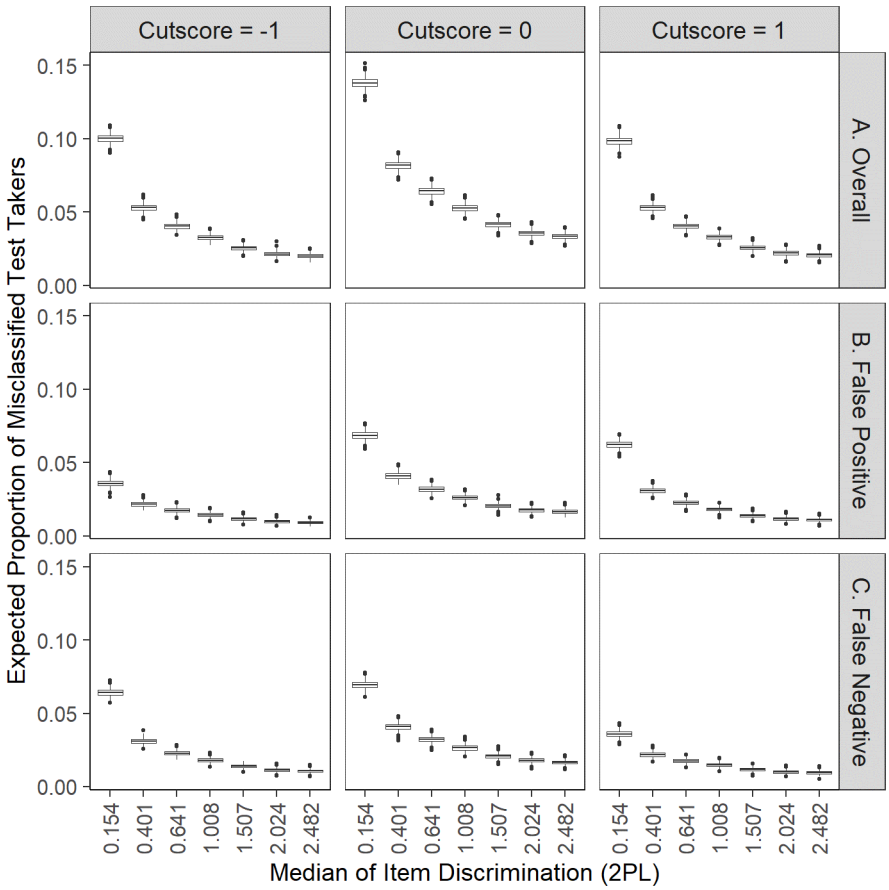

Figure 2
Expected proportion of misclassified test takers as a function of item discrimination for simulations with one cut score using 2PL IRT model.

Note. Data points represent outliers within each condition.

level 3 when estimated ability score belongs to performance level 2, and a false performance level 1 and a false performance level 2 when estimated ability score belongs to performance level 3. The remaining rows of Figure 3 show these six expected misclassification values.

The second row of Figure 3 shows that item discrimination is associated to a decreasing expected false performance level 2 classification of test takers whose estimated ability score belongs to performance level 1, and this trend is comparable across the cut score values. On the other hand, the third row shows that it is unlikely for test takers to be misclassified into performance level 3 if their estimated ability score belongs to performance level 1, unless the second cut score is placed in the mean of ability distribution and item discrimination is as low as .25. Although this is consistent with the general trend observed until now, misclassification into performance level 3 is still unlikely.

The fourth and fifth rows of Figure 3 show that increasing item discrimination is again associated to a decreasing expected misclassification of test takers into performance levels 1 and 3 when their estimated ability score belongs to performance level 2. In the first case, the false performance level 1 classification is comparable across cut score values because the first cut score was always fixed, so the proportion of misclassi-

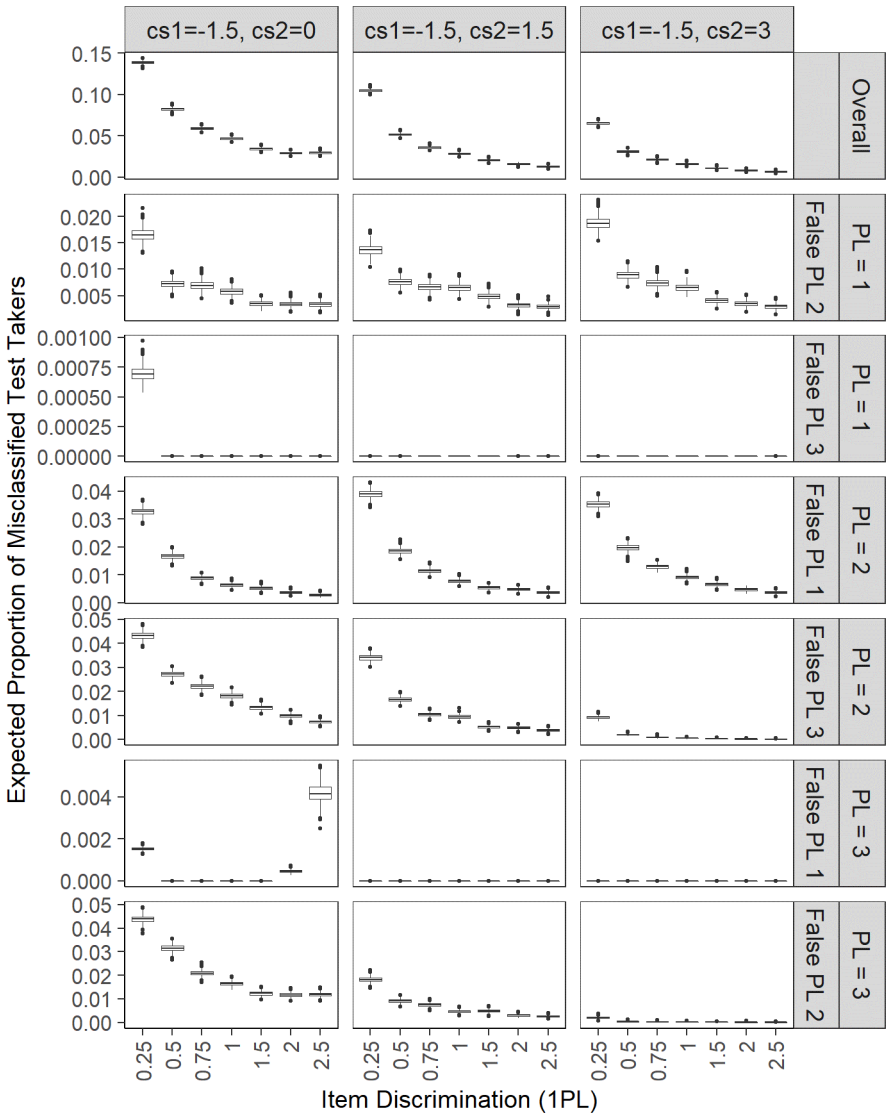

Figure 3

Expected proportion of misclassified test takers as a function of item discrimination for simulations with two cut scores using 1PL IRT model.

Note. Data points represent outliers within each condition. Notice the different scales in the *y*-axis. cs1 = cut score 1 (always fixed at -1.5); cs2 = cut score 2; PL = performance level.

fied test takers remained somewhat stable across simulations. In the second case, the false performance level 3 classification decreased as the second cut score shifted away from the mean of ability distribution, but misclassifications still increased as item discrimination decreased.

The last two rows of Figure 3 show the expected misclassification of test takers into performance levels 1 and 2 when their estimated ability score belongs to performance level 3. The sixth row shows that it is unlikely for test takers to be misclassified into performance level 1 if their estimated ability score belongs to performance level 3. An exception occurred when the second cut score is placed in the mean of ability distribution and item discrimination is 2.5, which is inconsistent with the general trend observed previously because one misclassification increased with a high

item discrimination. The reason for this exception is that TIF associated to an item discrimination of 2.5 is lower than TIF associated to other item discrimination values (data not shown) for ability values of 1 or higher, that is, at performance level 3; this lower information increased the standard error of measurement more than for other item discrimination values and increased test takers' expected classification inaccuracy. The last row of Figure 3 shows that the false performance level 2 classification decreased as the second cut score shifted away from the mean of ability distribution, but once again, misclassifications still increased as item discrimination decreased.

Finally, Figure 4 shows boxplots of the expected proportion of misclassified test takers as a function of item discrimination for simulations with two cut scores using 2PL IRT model. Notice again the different scales on the *y*-axis.

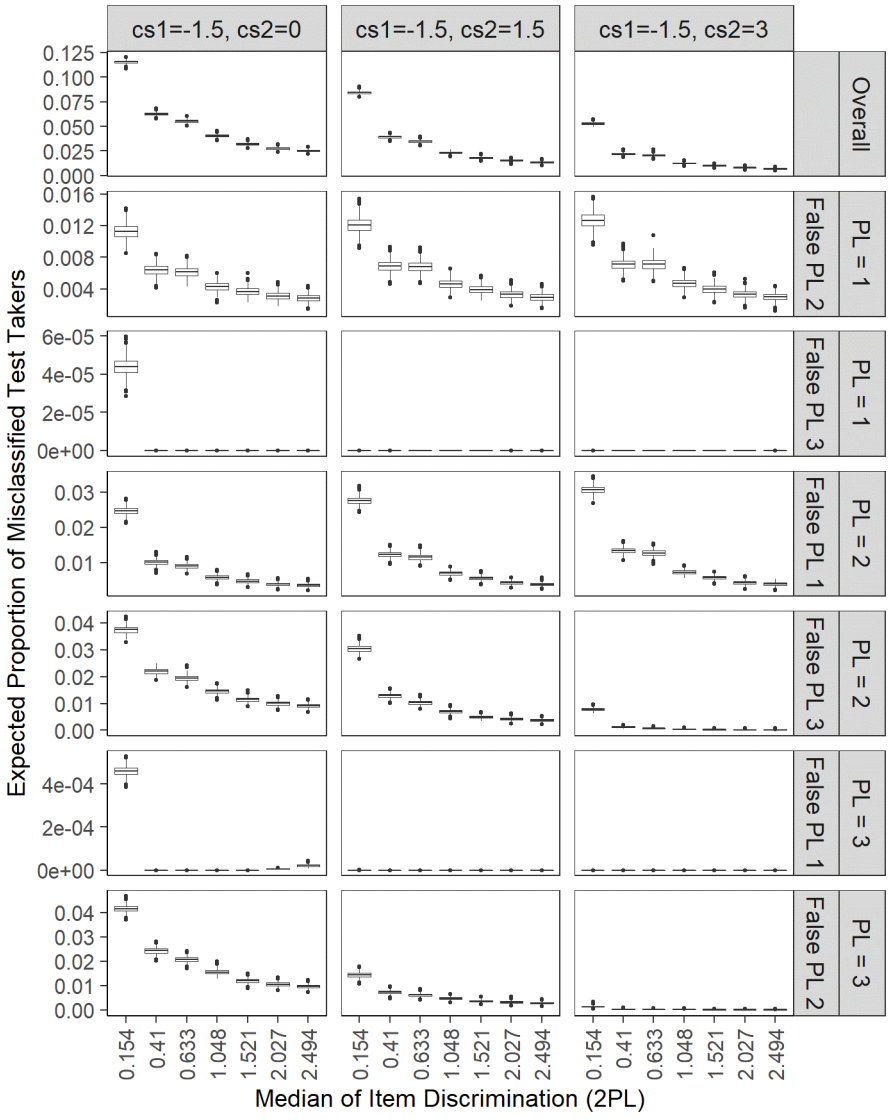

Figure 4
Expected proportion of misclassified test takers as a function of item discrimination for simulations with two cut scores using 2PL IRT model.

Note. Data points represent outliers within each condition. Notice the different scales in the *y*-axis. cs1 = cut score 1 (always fixed at -1.5); cs2 = cut score 2; PL = performance level.

Results using 2PL model are similar to those using 1PL model; if anything, the exception found using 1PL model when the second cut score is placed in the mean of ability distribution and item discrimination is 2.5 was not reproduced using 2PL model.

When Figures 3 and 4 are compared, the proportions of misclassified test takers are lower for the 2PL than for the 1PL model. Compare, for example, the top panels of both figures: the maximum overall expected misclassifications reach a median of approximately .112 for 2PL when the second cut score equals 0, lower than the median of approximately .140 for 1PL.

Discussion

The present simulations were conducted in order to test whether item discrimination influences the expected proportion of misclassified test takers. Consistent with previous studies (Lathrop & Cheng, 2013; Luecht, 2016; Xing & Hambleton, 2004), the results suggest this is the case: a test with low discriminating items tends to increase the proportion of misclassified test takers, irrespective of the location of the cut score. Only one exception was observed: Misclassifications into performance level 1 of test taker who should be placed into performance level 3 were higher when two cut scores were simulated, the second cut score was placed in the mean of ability distribution and item discrimination using the 1PL model was 2.5 (see Figure 3). However, this result does not undermine the general conclusion because this increase in misclassifications is associated to a lower TIF for ability values of 1 or higher; since item discrimination and the observed exception are both associated to TIF, which at the same time is associated to a correct -or incorrect- classification of test takers, then the general result and the observed exception do not contradict each other because both support that TIF is associated to classification inaccuracy.

In addition, simulation using the 2PL IRT model yielded less expected classification inaccuracies than simulation using the 1PL model. This may be attributed to the variability in discrimination values simulated under the 2PL model: Even when the median was as low as .154, there was at least one item with high discrimination (see the last column of Tables 1 and 2), thus increasing TIF and, therefore, reducing misclassifications.

The present results also replicated those of previous research suggesting that the number of cut scores and their location relative to the test takers ability distribution influence classification inaccuracy (Ercikan & Julian, 2002; Lathrop & Cheng, 2013; Lee, 2010; Martineau, 2007; Wyse & Hao, 2012). Specifically, misclassifications tended to decrease as the cut score value shifted away from the mean of the ability distribution, and this trend was more notorious when item discriminations (1PL) or their median (2PL) were low. Besides, with one cut score, false positives and negatives follow a different trend depending on the cut score value: the former decreased only when cut score equals -1, whereas the latter decreased only when cut score equals 1. This may have some implication depending on the cost associated to each misclassification (for example, when establishing a cut score to detect intellectual risk; see Ramírez-Benítez, Jiménez-Morales, & Díaz-Bringas, 2015), and minimization of one at the expense of the other may be carefully considered on each particular case. But still, item discrimination may help diminish this problem.

With two cut scores, misclassification into an adjacent performance level was higher than misclassification into a further one. Test takers classified into performance level 1 were more

likely to be misclassified into performance level 2 than into performance level 3; conversely, test takers classified into performance level 3 were more likely to be misclassified into performance level 2 than into performance level 1. In addition, test takers classified into performance level 2 were the most likely to be misclassified into any other performance level than the remaining test takers, but as a consequence of fixing the first cut score at -1.5, they were less likely to be misclassified into performance level 3 as the second cut score shifted away from the mean of the ability distribution, and this shift did not influence misclassification into performance level 1. But once again, item discrimination may help diminish this problem irrespective of the location of cut score values.

The manipulation of the second cut score simulated in the present study may not be far from some real-life situations. As an example, suppose that test takers who get scores at performance level 3 are candidate for an award; if the stakeholders decide to increase the minimum score to be classified into that level, some test takers who could have been eligible for receiving the award will no longer be considered because their test score will now belong to performance level 2. However, this decision will reduce the expected proportion of misclassified test takers in both performance levels. Every particular case should weigh the importance of each consequence in order to make a decision, but once again, increasing item discrimination may help stakeholders in decision making since misclassifications would be a less complicated issue to weigh.

Ercikan and Julian (2002) and Martineau (2007) further showed that the expected classification inaccuracy increases as the number of classification categories increases. Although the present study did not simulate conditions that are fully comparable between one and two cut scores, an exercise can be made only for illustrative purposes: comparing the overall expected proportion of misclassified test takers in conditions where the single cut score equals 0 versus conditions with two cut scores where the second cut score equals 0. This means comparing the top middle panel of Figure 1 with the top left panel of Figure 3, as well as the top middle panel of Figure 2 with the top left panel of Figure 4. Table 3 shows these comparisons more directly by reproducing the median misclassification values of the conditions just mentioned, their interquartile deviation (which is [P75 – P25]/2) and the difference between the medians. Positive differences imply higher misclassification with one cut score, negative differences imply higher misclassification with two cut scores.

As can be seen, all differences suggest that misclassifications were higher with one cut score that with two, which is contrary to the two studies previously mentioned. However, the differences are negligible and they decrease as item discrimination increases. One possible reason for this apparent discrepancy is the number of items simulated: the present study simulated responses to 100 items, whereas the previous studies simulated less than 60, so the present study simulated conditions in which negligible differences could be found since more items tended to increase TIF and, therefore, reduce the standard error of measurement. But once again, these results should be taken only as an illustrative exercise since comparability between conditions is not warranted.

Two limitations of the present results need to be considered. First, this study used the Rudner (2001, 2005) algorithm for estimating expected misclassification of test takers, which assumes that an IRT model fit data appropriately and that ability is normally distributed. This implies that the present results may not be generalizable to cases where these assumptions do not hold. In that

Table 3
Median (and interquartile deviation) of the overall expected proportion of misclassified test takers, and their difference, as a function of either item discrimination (1PL) or median of item discrimination (2PL).

	Expected misclassifications by number of cut scores		
Discrimination	**One (cs = 0)**	**Two (cs2 = 0)**	**Difference**
1PL			
0.250	.179 (0.0024)	.139 (0.0012)	.040
0.500	.116 (0.0022)	.083 (0.0011)	.033
0.750	.085 (0.0020)	.059 (0.0011)	.026
1.000	.067 (0.0018)	.047 (0.0010)	.020
1.500	.049 (0.0016)	.034 (0.0010)	.015
2.000	.040 (0.0015)	.029 (0.0009)	.011
2.500	.034 (0.0014)	.030 (0.0009)	.004
2PL			
0.154	.138 (0.0023)	.116 (0.0011)	.022
0.410	.082 (0.0020)	.063 (0.0010)	.019
0.633	.064 (0.0019)	.055 (0.0010)	.009
1.048	.053 (0.0016)	.040 (0.0010)	.013
1.521	.042 (0.0015)	.032 (0.0009)	.010
2.027	.036 (0.0015)	.028 (0.0008)	.008
2.494	.033 (0.0014)	.025 (0.0008)	.008

case, a nonparametric algorithm such as Lathrop and Cheng's (2014) may be more appropriate to estimate classification inaccuracy, but it is difficult to say whether item discrimination influences inaccuracy since no explicit relation has been stated between these two in an algorithm like that: the probability of a correct response is conditional on observed total score and these two are not related by an item characteristic curve determined by item parameters.

Another limitation is that item parameters were not estimated, instead, the simulated true values were used in estimation of the expected proportion of misclassified test takers, so capitalization on chance was not investigated. Hambleton and Jones (1994) mentioned that item parameter estimates have a positive error relative to their true values, and Yen (1987) showed that this error is bigger for discrimination parameter estimates. For the present study, this implies that TIFs could have been overestimated, and thus misclassifications in general would have decreased, had calibrated item parameter estimates been used. However, there is no reason to suspect that capitalization on chance has a differential influence on item discrimination depending on their true values, so it is possible that the main result of the present study may remain using parameter estimates instead of true values. In addition, Hambleton and Jones (1994) found that a sample size of 2000 test takers, such as the one used in the present study, reduces the effect of capitalization on chance, and Yen (1987) reported a diminishing effect of capitalization on chance as test length increased (see her Table 3), so it is reasonable to suspect that the same would have happened in this study, had calibrated item parameter estimates been used.

In summary, this study found that item discrimination has a negative association with the expected proportion of misclassified test takers: the higher the item discrimination becomes, the lower expected misclassification will be observed. In a test with a criterion-referenced score interpretation, it is important to get validity evidence based on test content (Popham & Husek, 1969), which is the reason that justifies the inclusion of item that don't fully discriminate (Burton, 2001; Clifford, 2016; Frisbie, 2005; Haladyna, 2016; Popham & Husek, 1969). Nevertheless, it is recommended to include as few items with low discrimination values as possible —or even none— because, otherwise, it becomes more likely to classify a test taker into a wrong performance level.

References

Baker, F. B., & Kim, S.-H. (2017). *The basics of Item Response Theory using R.* New York, N.Y: Springer. doi: 10.1007/978-3-319-54205-8

Burton, R. F. (2001). Do item-discrimination indices really help us to improve our tests? *Assessment & Evaluation in Higher Education, 26*(3), 213-220. doi: 10.1080/02602930120052378

Cheng, Y., Liu, C., & Behrens, J. (2015). Standard error of ability estimates and the classification accuracy and consistency of binary decisions. *Psychometrika, 80*(3), 645-664. doi: 10.1007/s11336-014-9407-z

Clifford, R. (2016). A rationale for criterion-referenced proficiency testing. *Foreign Language Annals, 49*(2), 224-234. doi: 10.1111/flan.12201

DeMars, C. (2010). *Item Response Theory.* Oxford, Oxfordshire: Oxford University Press.

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15*(3), 269-294. doi: 10.1207/S15324818AME1503_3

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24*(3), 21-28. doi: 10.1111/j.1745-3992.2005.00016.x

Haladyna, T. M. (2016). Item analysis for selected-response test items. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed, pp. 392-409). New York, NY: Routledge.

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are multiple-choice items too fat? *Applied Measurement in Education, 32*(4), 350-364. doi: 10.1080/08957347.2019.1660348

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7*(3), 171-186. doi: 10.1207/s15324818ame0703_1

Lathrop, Q. N. (2014). R package cacIRT: Estimation of classification accuracy and consistency under item response theory. *Applied Psychological Measurement, 38*(7), 581-582. doi: 10.1177/0146621614536465

Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R package cacIRT. *Practical Assessment, Research, and Evaluation, 20*, Article 18. Retrieved from https://scholarworks.umass.edu/pare/vol20/iss1/18

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. doi: 10.1177/0146621612471888

Lathrop, Q. N., & Cheng, Y. (2014). A nonparametric approach to estimate classification accuracy and consistency. *Journal of Educational Measurement, 51*(3), 318-334. doi: 10.1111/jedm.12048

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1-17. doi: 10.1111/j.1745-3984.2009.00096.x

Leydold, J., & H"ormann, W. (2021). Runuran: R interface to the 'UNU.RAN' random variate generators (Version 0.34) [R package]. Retrieved from https://

cran.r-project.org/web/packages/Runuran/index.html

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (2016). Applications of item response theory: Item and test information functions for designing and building mastery tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (2nd ed.). New York, NY: Routledge.

Martineau, J. A. (2007). An expansion and practical evaluation of expected classification accuracy. *Applied Psychological Measurement, 31*(3), 181-194. doi: 10.1177/0146621606291557

Paek, I., & Han, K. T. (2013). IRTPRO 2.1 for Windows (item response theory for patient-reported outcomes). *Applied Psychological Measurement, 37*(3), 242-252. doi: 10.1177/0146621612468223

Partchev, I., Maris, G., & Hattori, T. (2017). irtoys: A collection of functions related to item response theory (IRT) (Version 0.2.1) [R package]. Retrieved from https://cran.r-project.org/package=irtoys

Popham, W. J. (2014). Criterion-referenced measurement: Half a century wasted? *Educational Leadership, 71*(6), 62-66. Retrieved from http://www.ascd.org/publications/educational_leadership/mar14/vol71/num06/Criterion-Referenced_Measurement@_Half_a_Century_Wasted%C2%A2.aspx

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1-9. doi: 10.1111/j.1745-3984.1969.tb00654.x

R Core Team. (2020). R: A language and environment for statistical computing (Version 4.0.2). [Computer software]. Retrieved from https://www.R-project.org

Ramírez-Benítez, Y., Jiménez-Morales, R. M., & Díaz-Bringas, M. (2015). Matrices progresivas de Raven: Punto de corte para preescolares 4 - 6 años. *Revista Evaluar, 15*(1), 123-133. doi: 10.35670/1667-4545.v15.n1.14911

Richaud de Minzi, M. C. (2008). Nuevas tendencias en psicometría. *Revista Evaluar, 8*(1), 1-19. doi: 10.35670/1667-4545.v8.n1.501

Rizopoulos, D. (2018). ltm: Latent trait models under IRT (Version 1.1-1) [R package]. Retrieved from https://CRAN.R-project.org/package=ltm

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research, and Evaluation, 7*, Article 14. doi: 10.7275/an9m-2035

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research, and Evaluation, 10*, Article 13. doi: 10.7275/56a5-6b14

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). New York, NY: Springer. doi: 10.1007/978-3-319-24277-4

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., ... & RStudio. (2021). ggplot2: Create elegant data visualisations using the grammar of graphics (Version 3.3.5) [R package]. Retrieved from https://cran.r-project.org/web/packages/ggplot2/index.html

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement, 36*(7), 602-624. doi: 10.1177/0146621612451522

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*(1), 5-21. doi: 10.1177/0013164403258393

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*(2), 275-291. doi: 10.1007/BF02294241

Appendix 1. R Code Used to Conduct Simulations with One Cut Score

This appendix shows the R code I used to simulate each condition with one cut score. It is important to point out that after step 3a of the code I manually paired item discriminations with item difficulty values in such a way as to maximize test information around the cut score. Once done, the rest of code can be run.

```
#Simulate expected misclassifications with IRT 1PL or 2PL models
#One cut score

#Load necessary R packages
library(tidyverse)
library(irtoys)
library(Runuran)

#Set constants
nsimul <- 1000    #Number of replications of steps 4 to 9 according to step 10
n <- 2000            #Number of test takers according to step 4
k <- 100             #Number of items

#Steps 1 and 2: Set the number and value of cut scores
cutscore <- 0
#Step 3: Set the item parameter values
#3a: Sampling or setting item parameters
#Sampling item difficulty values
b <- rnorm(k,0,1)
#Fixing guessing parameter values at zero
c <- rep(0,k)
#When using 1PL model, set discrimination with these two lines
a1pl <- 0.25       #Item discrimination value
a <- rep(a1pl,k)
#When using 2PL model, sample discrimination with these two lines
meanlog <- -2    #Mean of sampled lognormal distribution
a <- urlnorm(k, meanlog, 1, 0, 3)
#3b: Creating table of item parameter values
#(column 1=a, column 2=b, column 3=c)
params <- cbind(a,b,c)
```

```
params <- data.matrix(params)

#Loop to perform 1000 replications
Data <- rep(NA,nsimul*3)
for (s in 1:nsimul) {
   #Step 4: Draw a sample of 2000 test takers from a
   #            standard normal distribution (~N(m=0,sd=1))
   theta_sim <- rnorm(n,0,1)
   #Step 5: Simulate 100 responses to dichotomously scored test items
   resps <- sim(params,theta_sim)
   #Step 6: Estimate test takers' maximum likelihood ability
   #            and standard error of measurement
   theta_obs <- mlebme(resps,params,method="ML")

   #Step 7: Estimate individual probability of misclassification
   inacc <- matrix(NA,n,2)
   for (i in 1:length(theta_sim)) {
      if (theta_obs[i,1]<cutscore) {
         #If ability < cutscore, label and estimate false positive
         inacc[i,1] <- 0
         inacc[i,2] <- pnorm(Inf,theta_obs[i,1],theta_obs[i,2]) -
                              pnorm(cutscore,theta_obs[i,1],theta_obs[i,2])
         }
      else {
         #If ability >= cutscore, label and estimate false negative
         inacc[i,1] <- 1
         inacc[i,2] <- pnorm(cutscore,theta_obs[i,1],theta_obs[i,2])
      }
   }

   #Step 8: Estimate overall expected misclassifications
   inacc <- data.frame(inacc_type=inacc[,1],value=inacc[,2])
   Data[s] <- mean(inacc$value)

   #Step 9: Estimate the expected proportion of false positives
   #            and false negatives
```

```
    #9a. False positives
    falsepos <- inacc %>% filter(inacc_type==0)
    Data[s+nsimul] <- mean(falsepos$value)*(length(falsepos$value)/n)
    #9b. False negatives
    falseneg <- inacc %>% filter(inacc_type==1)
    Data[s+nsimul*2] <- mean(falseneg$value)*(length(falseneg$value)/n)

    #Remove objects from R workspace
    rm(falsepos,falseneg,inacc,resps,theta_obs,theta_sim)
}
```

Appendix 2. R Code Used to Conduct Simulations With two Cut Scores

This appendix shows the R code I used to simulate each condition with two cut scores. Once again, after step 3a of the code I manually paired item discriminations with item difficulty values in such a way as to maximize test information around the cut scores. Once done, the rest of code can be run.

```
#Simulate expected misclassifications with IRT 1PL or 2PL models
#Two cut scores

#Load necessary R packages
library(tidyverse)
library(irtoys)
library(Runuran)

#Set constants
nsimul <- 1000 #Number of replications of steps 4 to 9 according to #step 10
n <- 2000    #Number of test takers according to step 4
k <- 100    #Number of items

#Steps 1 and 2: Set the number and value of cut scores
cutscore1 <- -1.5
cutscore2 <- 0
#Step 3: Set the item parameter values
#3a: Sampling or setting item parameters
#Sampling item difficulty values
b_cs1 <- urnorm(50, cutscore1, 0.5, cutscore1-0.75, cutscore1+0.75)
b_cs2 <- urnorm(50, cutscore2, 0.5, cutscore2-0.75, cutscore2+0.75)
b <- c(b_cs1,b_cs2)
#Fixing guessing parameter values at zero
c <- rep(0,k)
#When using 1PL model, set discrimination with these two lines
a1pl <- 0.25   #Item discrimination value
```

```
a <- rep(a1pl,k)
#When using 2PL model, sample discrimination with these five lines
meanlog_cs1 <- -2   #Mean of sampled lognormal distribution
meanlog_cs2 <- -2   #Mean of sampled lognormal distribution
a_cs1 <- urlnorm(50, meanlog_cs1, 1, 0, 3)
a_cs2 <- urlnorm(50, meanlog_cs2, 1, 0, 3)
a <- c(a_cs1,a_cs2)
#3b: Creating table of item parameter values
#(column 1=a, column 2=b, column 3=c)
params <- cbind(a,b,c)
params <- data.matrix(params)

#Loop to perform 1000 replications
Data <- rep(NA,nsimul*7)
for (s in 1:nsimul) {
 #Step 4: Draw a sample of 2000 test takers from a
 #    standard normal distribution (~N(m=0,sd=1))
 theta_sim <- rnorm(n,0,1)
 #Step 5: Simulate 100 responses to dichotomously scored test items
 resps <- sim(params,theta_sim)
 #Step 6: Estimate test takers' maximum likelihood ability
 #    and standard error of measurement
 theta_obs <- mlebme(resps,params,method="ML")

 #Step 7: Estimate individual probability of misclassification
 inacc <- matrix(NA,n,3)
 for (i in 1:length(theta_sim)) {
  if (theta_obs[i,1]<cutscore1) {
   #If ability < cutscore1, label at performance level 1
   inacc[i,1] <- 1
   inacc[i,2] <- pnorm(cutscore2,theta_obs[i,1],theta_obs[i,2]) -
          pnorm(cutscore1,theta_obs[i,1],theta_obs[i,2])
   inacc[i,3] <- pnorm(Inf,theta_obs[i,1],theta_obs[i,2]) -
          pnorm(cutscore2,theta_obs[i,1],theta_obs[i,2])
   }
  else if (theta_obs[i,1]>=cutscore2) {
   #If ability >= cutscore2, label at performance level 3
   inacc[i,1] <- 3
   inacc[i,2] <- pnorm(cutscore1,theta_obs[i,1],theta_obs[i,2])
   inacc[i,3] <- pnorm(cutscore2,theta_obs[i,1],theta_obs[i,2]) -
          pnorm(cutscore1,theta_obs[i,1],theta_obs[i,2])
  }
  else {
   #If cutscore1 <= ability < cutscore2, label at performance level 2
   inacc[i,1] <- 2
   inacc[i,2] <- pnorm(cutscore1,theta_obs[i,1],theta_obs[i,2])
   inacc[i,3] <- pnorm(Inf,theta_obs[i,1],theta_obs[i,2]) -
          pnorm(cutscore2,theta_obs[i,1],theta_obs[i,2])
  }
 }

 #Step 8: Estimate overall expected misclassifications
 inacc <- data.frame(perf_level=inacc[,1],lower=inacc[,2],upper=inacc[,3])
 Data[s] <- mean(c(inacc$lower,inacc$upper))
```

```
 #Step 9: Estimate expected misclassifications
 #    at each performance level
 #9a. Misclassifications at performance level 1 (PL 1)
 pl1 <- inacc %>% filter(perf_level==1)
 #p(2|1) = False PL 2
 Data[s+nsimul] <- mean(pl1$lower)*((0.5*length(pl1$lower))/n)
 #p(3|1) = False PL 3
 Data[s+nsimul*2] <- mean(pl1$upper)*((0.5*length(pl1$upper))/n)

 #9b. Misclassifications at performance level 2 (PL 2)
 pl2 <- inacc %>% filter(perf_level==2)
 #p(1|2) = False PL 1
 Data[s+nsimul*3] <- mean(pl2$lower)*((0.5*length(pl2$lower))/n)
 #p(3|2) = False PL 3
 Data[s+nsimul*4] <- mean(pl2$upper)*((0.5*length(pl2$upper))/n)

 #9c. Misclassifications at performance level 3 (PL 3)
 pl3 <- inacc %>% filter(perf_level==3)
 #p(1|3) = False PL 1
 Data[s+nsimul*5] <- mean(pl3$lower)*((0.5*length(pl3$lower))/n)
 #p(2|3) = False PL 2
 Data[s+nsimul*6] <- mean(pl3$upper)*((0.5*length(pl3$upper))/n)

 #Remove objects from R workspace
 rm(pl1,pl2,pl3,inacc,resps,theta_obs,theta_sim)
}
```