Chapter Six: Using Binomial Logistic Regression and Receiver Operating Curves to Identify at Risk Learners

Celeste Combrinck
Faculty of Education, University of Pretoria, South Africa

Introduction and background

The subjectivity of standard setting as well as the importance of the decisions to be taken based on these standards, which affect the future of the learners tested, are among the main reasons that make the setting of cut scores one of the most complex, contradictory and controversial problems in the area of achievement testing (Kaftandjieva 2010:10).

Standard and norm-setting, determining cut-off points and decision criteria can be done arbitrarily, with reasonable sounding points being selected or done scientifically and based on data for or from the instrument(s) (Blömeke and Gustafsson, 2017). Test users sometimes choose to set such criteria based on a perceived reasonable number, for example, 50 per cent or 60 per cent (Habibzadeh et al. 2016; Yousef et al. 2017). Such a number is chosen because the test users are unfamiliar with other methods for setting cut scores, and the number used has a tradition attached to it (Khatimin et al. 2013; Kubiszyn and Borich, 2024). Unfortunately, cut scores, even for high-stakes tests, are often set using a number that is not more precisely chosen. A failure to set evidence-based cut points has social justice implications and negatively affects scientific accuracy (Kellow and Wilson, 2008; Yudkowsky et al. 2015). Setting such points more scientifically requires technical knowledge and other resources, such as a panel of experts, to review and rate the test or items. This paper is a practical demonstration of how to set cut-off points scientifically using the Rasch partial credit model (Masters 2016). I demonstrate how to assess the validity assumptions related to instruments and convert the results to an interval scale using logistic regression modelling and receiver operating

curves (ROC). These methods use only data and are recommended when other resources, such as the judges or experts, are not available due to financial or practical constraints experienced by the test users (Diniz 2022).

Setting cut scores requires research studies involving several professionals with the appropriate backgrounds and consequently be expensive and time-consuming, as is typical of the most preferred Angoff method (Khalid et al. 2022). For these reasons, setting cut scores is often not practically possible in situations where it is required, such as in educational settings with limited resources (Cohen-Schotanus and Van der Vleuten 2010; Wang and Keller, 2024). In literature, criterion-referenced methods for setting cut scores are advocated as best practice (Park et al. 2021; Parsaeian et al. 2024;). However, this may not be possible for the test users to employ (Kellow and Wilson 2008; Pitoniak and Morgan 2017; Smith and Stone 2009). Therefore, the current manuscript explores non-parametric statistical techniques and reports on a purely data-driven approach, usefulness in setting cut scores.

Measurement plays a vital role in setting standards and cut-score determination while forming a part of an evaluative decision-making process (MacCann and Stanley 2006). Measurement theory and data inform the decision-making process, but should alone not dictate the outcome (Stone 2011). Cut scores should be linked to the construct and what the person is expected to have mastered, crucial aspects of the process which are lost when such points are set arbitrarily (Wyse 2017). Reducing the score could result in a loss of mastery of aspects of the content, as contained in specific items or scales in the instrument (Stone 2011). A reasonable cut score is directly related to the instrument's construct validity regarding the content coverage that should ideally be grasped. A test score is not necessarily equivalent to competency; mastery and competency differ (Gervais 2016). Instead, any test, no matter how well designed, should form a more comprehensive process, increasing the overall validity of decisions in which assessments play a role (Stone 2011). There is no universal method for setting standards and cut scores; instead, several different methods may apply to certain test types, and researchers and practitioners only sometimes have sufficient knowledge of these methods to use them (Wyse 2017). Testing and instrument development are also evolving, therefore,

methods for setting standards and cut scores are required to adapt and could be done using generative artificial intelligence (Latif and Zhai 2024). Furthermore, different methods can lead to various cut scores being set (Carlson et al. 2009; Kaftandjieva 2010).

Angoff's (1971) method has remained the most popular and well-known method for setting cut scores. It requires each item to be rated by a panel of experts in terms of whether a person of minimal proficiency would be able to answer the question correctly, and these items are then totalled to reach a cut score (Angoff 1971; Biddle 1993). This process requires the judges to understand what a minimally proficient person must know to succeed in a given scenario. A modification to this process could be to ask the raters what the probability is that a minimally proficient person or persons would correctly answer the item, in which case the sum of the probabilities would be used to calculate an acceptable cut score (Wyse 2017). The process requires at least seven to ten raters of the items (Biddle 1993). Most standard-setting methods involve using raters to assess whether a "proficient" person would correctly answer an item, depending on what such a person would be expected to know (Cizek et al. 2004; Kaftandjieva 2010). When Rasch person measures are linked to scales and further validated through convergent validity, using person measures to set standards becomes possible instead of relying on raters or judges (MacCann and Stanley 2006). Criticism of Angoff's methods is that judges need help to estimate item response probabilities for minimally competent candidates (Ricker 2006). Decision and overconfidence biases are also drawbacks of Angoff's method (Longford 1996). Such methods require not only a panel of experts, but also that such experts receive training to reduce their propensity to give biased ratings (Arce and Wang 2012).

Various authors have explored and advocated the combination of using judges and Rasch item measures as a more efficient way to set cut scores (Arce and Wang 2012; Baghaei 2007; Boone et al. 2014; Bowers and Shindoll, 1989; Wright, 2000). This method, however, also has challenges as judges may only sometimes agree with the Rasch measures regarding where the items are located, and a compromise must be reached (Baghaei 2007, 2009). A further challenge of using panel methods is that considerable time should be spent training panellists to minimise bias (Schultz 2006; Wyse

2018). Arce and Wang (2012) found that panellists can add significant increases to the standard error when setting cut scores and that such cut scores could vary according to the training and instructions given to panellists. The authors compared traditional and alternative approaches to using the modified Angoff's method and were unable to conclude whether this produced more accurate cut scores for academic achievement levels (Arce and Wang 2012). They also report that developing a standard setting and research framework is time-consuming. Different approaches when using panellists may result in very different outcomes, primarily due to the subjective nature of setting cut scores when using human opinions and ratings (Tannenbaum and Kannan 2015; Wang 2003). Khatimin et al. (2013) used the Rasch objective standard-setting method for setting standards and cut scores (Grosse and Wright 1987; Wright and Grosse 1993). Wright and Grosse (1993) also found that judges' ratings vary widely depending on factors such as expectations of the process, the scope presented and their specific expertise. According to Khatimin et al. (2013), using the Rasch Objective Standard Setting mitigates some problems with panellists to set cut scores and standards.

Data-based methods for setting cut scores and standards and identifying at-risk learners include predictive validity and statistical methods (Diniz 2022). Criterion and predictive validity of reading assessments have been used in the United States to predict how well learners achieve on high-stakes tests (Klingbeil et al. 2015). Curriculum-based assessment can broadly indicate knowledge and skills gained in a learning area (Mitchell 2019). Silberglitt and Hintze (2005) compared discriminant analysis, logistic regression and ROC curves to see which worked best for setting curriculum-based measurements. The authors found that each method could set adequate cut scores, which led to high specificity and negative predictive power, which they attribute to the fact that each method attempts to maximise the number of true negatives. Silberglitt and Hintze (2005) conclude that logistic regression is the most parsimonious method, though each method gives good results. A study which examined ROC, T-scores and the Rasch rating scale method (RSM) for setting cut scores found that the ROCs and RSM identified similar cut scores, whereas the T-scores led to lower, more conservative scores being set (DiStefano and

Morgan 2011). Stone et al. (2011) point out the importance of investigating the construct validity of an instrument before applying standard-setting methods. In their comparison of the Angoff method and objective standard setting (OSS), Stone et al. (2011) found that Angoff's method did not define a stable and valid construct. The importance of construct validity points to the fact that theories such as Rasch measurement theory should be applied to the instrument before cut scores are set, as was done in the current study.

Methods

The current study was part of a more significant endeavour to design tests to gauge curriculum knowledge gained over a year at a group of South African schools (Combrinck 2018). The schools were independent and part of a coalition with its curriculum, partly based on the South African Curriculum Assessment Policy Statements (CAPS) (Department of Basic Education 2024). The independent schools in the current study have longer school days than nationally recommended, smaller classes and give learners more individual attention. The learners came from impoverished environments and attended low-functioning primary schools. These independent schools aim to prepare learners for their end-of-year exams in the final school year so that they attain access to tertiary studies (Combrinck et al. 2016). The learners are second-language English speakers but have been in English Language medium schools since Grade 1. The funders wanted the developed instruments to be used as benchmarking tools (comparing schools) and accountability measures and to feed back into the school system to enhance teaching and learning (Combrinck et al. 2017). The instruments were also designed to monitor learning progression and identify learners requiring additional assistance (Combrinck et al. 2018). This chapter looks at a sub-component of the study, in which scores were examined and cut scores set, which could assist teachers in identifying learners who had a higher risk of not finishing school with access to tertiary studies. Such cut scores could then be used to identify at-risk learners going forward and resources put in place to assist academically at-risk learners.

Sample

At the end of Grade 11 all learners at the cluster of schools wrote the English Language, Mathematics and Natural Science tests. Not all learners have all three subjects; 384 wrote the English language test, 360 the mathematics test, and 309 wrote the natural science test. Two hundred ninety-six learners wrote all three of the tests (Valid N). The current study combined data from two cohorts of two years' worth. All the schools in the cluster were assessed, and the seven schools are seen as the total population, a sub-population in the school system. Most of the sample was female (78 per cent), as a girl-only school forms part of the cluster, and the other schools also have more female learners than male learners. The sample chosen for this study had written their final school year exam, and the schools provided the results. A retrospective analysis was done to set cut scores and then compare the success of these scores to the Grade 12 outcomes.

Instruments

Before cut scores can be set, an assumption about the quality of the instrument is taken for granted. Assessments designed for the current study followed strict guidelines for best test design (Boateng et al. 2018; Kline 2015; Wright and Stone 1979). The instruments were designed to be curriculum-based measurements (CBM), assessing the curriculum's broad goals and giving a more general indication of knowledge and skills gained throughout the year (Hintze and Silberglitt 2005). Curriculum-based measurements help enhance teaching and learning, improve curriculum implementation, monitor learning progression, evaluate learning programmes and identify academically at-risk learners (Costello et al. 2022; Deno, 2003; Hintze and Silberglitt 2005).

Subject specialists identified learning goals based on the national curriculum to assist with the test design (Kellow and Wilson 2008). After identifying the learning goals, the experts designed individual items to address the learning goals. The current study focussed on creating a range of items on a spectrum of difficulty and different levels of Bloom's taxonomy (Bloom et al. 1956). Next, the instruments were piloted and refined based

on Rasch statistics and subject specialist analysis of pilot results. Table 6.1 shows the Rasch criteria for quality instruments and to what degree each instrument.

Table 6.1: Instrument quality criteria

Criterion	English Test	Maths Test	Science Test	Interpretation
Targeting	0.08	0.13	0.07	Good targeting. Less than 1.0 error is good
Item Model Fit Mean Square Range	0.84–1.30	0.66–1.64	0.57–1.81	Productive for measurement
Person Reliability	0.86	0.86	0.81	Excellent
Item Reliability	0.98	0.97	0.95	Excellent
Ceiling Effect	None	None	None	Excellent
Floor Effect	None	None	None	Excellent
Variance in Data Explained by Measures (unidimensionality)	30.50%	34.80%	29.20%	Acceptable, no secondary dimensions (eigenvalues lower than 3.0)
Local Dependence	0.36–0.36	0.72–40	0.59–0.30	Values higher than 0.7 indicate dependence. Only one item in the Maths test exhibited potential dependence.

Criterion	English Test	Maths Test	Science Test	Interpretation
Sample Size	384 persons for 77 items	360 persons for 57 items	309 persons for 79 items	Acceptable sample size for dichotomous and polytomous items (99% confidence)
Person Separation Index	2.45	2.46	2.04	Values above 2 are desirable; all the tests had acceptable separation indexes
Item Separation Index	6.66	5.83	4.37	Values above 3 are desirable; all the tests had excellent separation indexes.

The English language, mathematics, and science assessments achieved most of the requirements, having excellent targets, model fit and person and item reliability, and they did not show any ceiling or floor effects. The criteria were based on recommendations from Linacre (2023b), Boone (2016), Bond et al. (2020) and Boone et al. (2014).

Processes

Data collection was done by the external monitoring agent each year in November, and all the Grade 11 learners (who were taking the respective subjects) wrote the three assessments. Experienced teachers scored the assessments, and scripts were moderated. Thereafter, all scripts were captured on the item level and the results were analysed. All procedures

were standardised. After these identical learners completed their twelfth year, their results were obtained from the schools for comparison with their Grade 11 results for predictive validity.

Data analysis

Two commonly used methods for setting cut scores in assessment measures were considered: binomial logistic regression and ROC curves. Each method has advantages and disadvantages; for example, ROC analysis allows setting different cut scores by comparing them to one another. Logistic regression uses maximum likelihood estimation to maximise the classification of true positives, but at the cost of less accurate identification of true negatives. Both methods were applied and compared to find the most accurate methods for setting the cut scores and identifying the academically at-risk learners.

The Rasch model

The Rasch model was applied to the instruments to assess their quality and to convert the ordinal scale to an interval scale. Rasch measurement theory is a family of statistical models, all of which are logistic regression models, in which the probability of correctly answering an item is calculated as the person's ability minus the item's difficulty (Combrinck 2020). The raw scores were transformed into an interval scale via the Rasch measurement model in Winsteps (Linacre 2023a). This had the advantage of giving each student a score based on the log odds of their correctly answering an item versus how many persons could correctly answer the item. The final person score was based on all items and person interactions, and the logit score was converted to a scale of 0 to 100, with the same mean as the raw scores and a standard deviation of 10.

Binomial logistic regression

To graduate from high school in South Africa, learners must pass the final 12th-year exams, that is, the National Senior Certificate (NSC). To pass the

exams, learners should have achieved 40 per cent or more in three subjects, and one of those subjects should be a language at the home language level (Umalusi 2014). The other two subjects require a minimum of 30 per cent or more. To gain access to diploma-level study, the learner must achieve 40 per cent or higher in their home language and 40 per cent or higher in at least three other subjects. To gain entry to bachelor studies, the learner must have 40 per cent or higher in five subjects, one of which must be the home language. As this coalition of schools aims to prepare learners to gain tertiary access, a dichotomous variable was created: no tertiary access failed or only passed, or access, diploma or bachelor access. Binomial logistic regression was used to assess whether the three tests could be used as predictors for the dichotomous outcomes. All assumptions were met for binomial logistic regression; variables were linearly related; there was a lack of multicollinearity and errors were independent (Field 2009; Field 2024). Linearity of the logit was assessed in SPSS (IBM 2025) using interaction terms, and all three of the interactions were not statistically significant ($p > 0.05$), indicating that the assumption of linearity of the logit holds (Field 2009). The potential multicollinearity problem was tested using collinearity statistics in the linear regression analysis option. The tolerance values were above 0.1, ranging from 0.419 to 0.679. The VIF values did not exceed 10; these ranged from 1.473 to 2.385; all these indicators point to multicollinearity not being a problem in this analysis (Myers 1990). Mathematics and Science scores correlate highly at 0.731 ($p = 0.000$), however, running the model without either weakens the model's predictive value. Forced entry was used as the subjects were expected to predict the outcome to some degree (Tabachnick and Fidell 2007). Two cases with high values for Cook's distance exceeded 1 and were removed from the analysis, reducing the sample size to 294. The standardised residual values were below 2 except for one case (less than 5 per cent of the sample), and the mean leverage value was close to the expected value of 0.010 at 0.013. The DFBetas had values of less than 1, and all residuals indicate that the model fits the data and that errors are independent.

ROCs

ROCs were used to gauge the discriminant value of tests for dichotomous outcomes and to determine the cut points for identifying at-risk learners (Metz 1978; Zweig and Campbell 1993). ROC results for mathematics and science test results were analysed separately, and cut scores were set discretely for these two subjects to maximise sensitivity and specificity. Sensitivity is the probability of the predictor correctly identifying true positives, in this case, not having tertiary access at the end of the schooling career (Grade 12). Specificity is the probability of the predictor correctly identifying true-negatives, in this case, those who would have tertiary access correctly being classified as such and not classified as being at risk (Linacre 1994; Metz 1978). The cut scores' usefulness was examined using a prediction table to see how many learners were correctly identified as at risk (true positive) and how many were not identified who should have been classified as such (false negative).

Results

Table 6.2 shows the descriptive statistics for the pass type regarding learners who had completed all three subjects. The end-of-school results showed that 172 (58 per cent) of the learners achieved tertiary access, passing either with a diploma or bachelor's degree. While 124 (42 per cent) had no tertiary access, they failed their exams or only passed.

Table 6.2: Descriptive of final year exam results for learners with mathematics and science

Outcome	Frequency	Per cent	Valid Percent	Cumulative Percent
Tertiary access	172	58.1	58.1	58.1
No tertiary access	124	41.9	41.9	100.0
Total	296	100.0	100.0	

In Table 6.3, the results of the binomial logistic regression are shown. The Grade 11 English language test did not significantly predict the dichotomous outcome (p = 0.792). For this reason, the Grade 11 mathematics and science tests were used to identify at-risk learners and the language test results were not included in the ROC analysis. The model fit statistics were insignificant; Hosmer and Lemeshow test = 3.629 (p = 0.898), indicating a good fit between the data and the model (Hosmer et al. 1997). The exponential statistic shows that for every percentage gained in the Grade 11 mathematics test, the learner becomes 1.590 times more likely to pass with tertiary access, and for the Grade 11 science test, for every extra percentage, the learner becomes 2.521 times more likely to pass with tertiary access. The pseudo-R2 is high, with 91.9 per cent of the variance in the outcome being explained by the model (Nagelkerke).

Table 6.3: Binomial logistic regression output for three test results predicting the type of pass

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
English Language Test results*	-0.038	0.039	0.973	1	0.324	0.963	0.892	1.038
Mathematics Test results*	0.464	0.094	24.347	1	0.000	1.590	1.322	1.911
Science Test results*	0.925	0.180	26.336	1	0.000	2.521	1.771	3.589
Constant	-47.119	8.834	28.453	1	0.000	0.000		

Note Pseudo R^2 = 0.919 (Nagelkerke), Hosmer and Lemeshow = 3.629 (p = 0.898), overall prediction = 96.3

* Note rescaled using Rasch model, 0–100 scale with STD of 10

Table 6.4 is the classification table and shows that, overall, the model had a prediction rate of 94.9%. The model correctly identified 95.6% of learners as having tertiary access and correctly identified 93.5% as not having tertiary access. The null model had an overall prediction rate of 58.1% and could

not identify any at-risk learners (0% correctly identified for this group). The result indicates that the Grade 11 Mathematics and Science tests significantly increase the possibility of identifying learners who will gain tertiary access at the end of Grade 12.

Table 6.4: Classification table

Observed	Predicted		
	Tertiary Access	No Tertiary Access	Percentage Correct
Tertiary access	167	5	97.1
No tertiary access	6	116	95.1
Overall Percentage			96.3

a. The cut value is 0.500

Table 6.5 presents the results using the predicted categories from the binomial logistic regression analysis to set cut scores. These cut scores were based on the mean of the Grade 11 tests for the predicted categories. The cut point was set for mathematics at 39.85 and science at 24.39 based on the predicted groups from the binomial logistic regression analysis. The results show that there were no false negatives, that is, those who would have tertiary access were identified correctly. However, there were false positives, such as learners who were at risk, but were not identified. Eighty-four learners were not identified as being at risk when the mathematics cut point was applied, and 68 learners were not identified when the science cut point was applied.

Table 6.5: Prediction table with binomial logistic regression categorisation

		No Tertiary Access	Tertiary Access	TOTAL
Maths cut point from logit rescaled applied	**At risk**	79	0	79
	Not at risk	84	197	281
	TOTAL	163	197	360
Science cut point from logit rescaled applied	**At risk**	61	0	61
	Not at risk	68	180	248
	TOTAL	129	180	309

Table 6.6 shows the results of the ROC analysis for the mathematics test. The area under the curve was 0.814, indicating a balance of sensitivity and specificity, a reasonably helpful test (Camasso and Jagannathan 1995). The statistically significant result (asymptotic = 0.000) provides further evidence of the instrument's accuracy for predicting a dichotomous outcome.

Table 6.6: Area under the curve for mathematics rescaled test

Area	Std. Error	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.814	0.026	0.000	0.763	0.865

a. Under the non-parametric assumption

b. Null hypothesis: true area = 0.5

The ROC plot for the mathematics test is presented below in Figure 6.1.

An intercept of 0.726 and 0.715 was used to maximise sensitivity and specificity. The modelling helped to identify these intercepts, resulting in a cut score of 43.920 for the mathematics results.

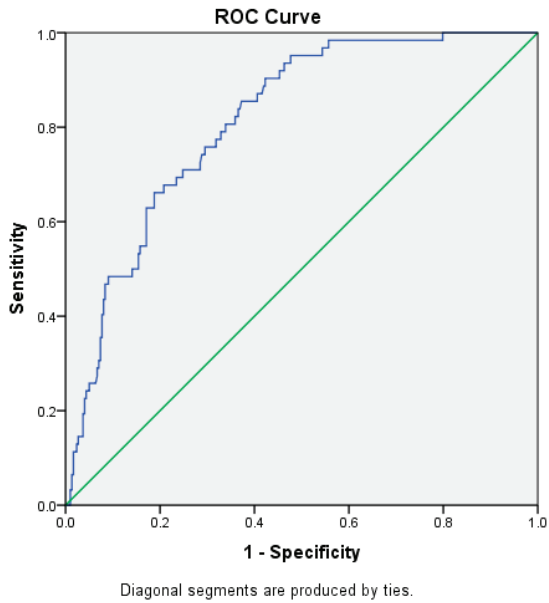

Figure 6.1: ROC curve for mathematics test data and dichotomous end-of-year outcome

Table 6.7 shows the results for the science test, with an area under the curve of 0.815, like that of the mathematics test and statistically significant (p = 0.000).

Table 6.7: Area under the curve for science rescaled test

Area	Std. Error	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.815	0.030	0.000	0.757	0.873

a. Under the non-parametric assumption

b. Under the non-parametric assumption

Table 6.7 reveals the ROC curve for the science test data. The aim was to identify a good trade-off between sensitivity and specificity, and an intercept of 0.754 and 0.728 was identified, which resulted in a cut score of 28.505 for the science results.

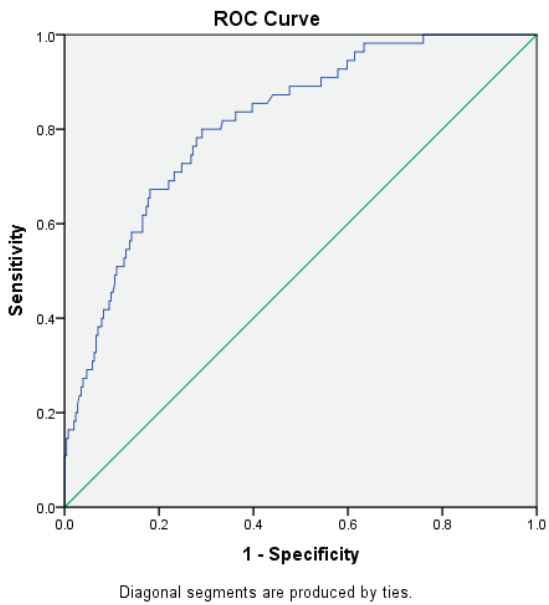

Figure 6.2: ROC curve for science test data and dichotomous end-of-year outcome

Based on these cut points, learners were classified as "at risk" or "not at risk" in Table 6.8 below and cross-tabulated with their results regarding whether they gained tertiary access or did not gain access. Based on the mathematics dichotomy, the categorisation correctly identified 45 learners as at risk while failing to identify seventeen who were at risk. For the science categorisation, 41 were correctly identified, and fourteen, also at risk, were not identified.

Table 6.8: Prediction table with ROC categorisation

		No tertiary access	Tertiary access	Total
Maths cut point from logit rescaled applied	**At risk**	45	85	130
	Not at risk	17	213	230
	Total	62	298	360
Science cut point from logit rescaled applied	**At risk**	41	69	110
	Not at risk	14	185	199
	Total	55	254	309

Combining the results from the mathematics classification and the science classification, the results are presented in Table 6.9. The combination of the two subjects leads to more learners being correctly identified.

Table 6.9: Classification table with ROC categorisation combined for two subjects

	No Tertiary Access	Tertiary Access	Total
At risk	45	79	124
Not at risk	7	165	172
Total	52	244	296

After combining the mathematics test and the science test cut-off scores, 45 learners were correctly identified as at-risk, and seven were not identified. In total, 79 were also identified as being at risk, but ultimately were not at risk as they obtained tertiary access. In practice, this would mean that out of 296, 124 learners would receive additional assistance to increase their chances of gaining tertiary access.

Discussion

Table 6.10 presents the results of the percentage correctly predicted by the binomial logistic regression analysis when cut scores were set for

mathematics and science separately, as well as the correctly predicted percentage when the model classified learners based on results.

Table 6.10: Classification table of percentages correctly predicted

Type	Observed			Predicted		
	Binomial Logistic Regression			Receiver Operating Curves		
	Maths	Science	Combined*	Maths	Science	Combined*
Tertiary access	100.0	100.0	97.1	92.6	93.0	95.9
No tertiary access	70.1	72.6	95.1	34.6	37.3	63.7
Overall Percentage	85.1	86.3	96.3	63.6	65.2	79.8

*Mathematics and science results combined for prediction

The following columns show the same results when ROCs are used to set the cut scores. Both logistic regression and ROC analysis are highly accurate at identifying those not at risk, yet, the logistic regression model is recommended as it is more accurate compared to the ROC. The most precise way to identify learners who require intervention is to use the predicted categories from the logistic regression model, which correctly identifies learners as being at risk 95 per cent of the time. Using logistic regression to set cut scores for mathematics and science separately results in 70 per cent and 73 per cent of learners being correctly identified. Having more measurement points increases the accuracy of correctly identifying the learners in danger of not achieving university allowance. It should also be noted that the time between the end of school results and the assessment can reduce the accuracy of the prediction; in this case, a full year had elapsed between the two measurement points, but the Grade 11 test results yielded accurate indications for Grade 12.

Implications for practice and policy

Like many resource-constrained countries, South Africa has limited access to cultural capital through measurement experts, psychometrists and assessment designers (Combrinck, 2018; Laher and Cockcroft 2017). There is a need to find alternative, more accessible techniques for setting cut scores and maintaining accountability indicators involving stakeholders (Kohrt and Kaiser 2021; Spaull 2015).

Accurate, valuable cuts-scores have important implications for human-centred decision-making which can be summarised as:

- Identifying students who require academic help.
- Informing policy in terms of how to set reliable cut scores.
- Setting and maintaining standards in educational decision-making.
- Enhancing the quality of national and school-based assessments so that accurate and valid inferences can be made for immediate and long-term use.

Areas where the current study can be applied include national assessments, early warning systems, benchmarking and performance indicators and resource allocation. For more sustainable futures in developing countries, one must consider the complexity of resource restraints and the desire to support students who need academic assistance the most. Using statistical techniques to find human-centred solutions can be done as demonstrated in the current study and applied in African contexts and other developing environments.

Implications for educational practice include the following poignant points:

Interventions for at-risk students: Predictive models can pinpoint the individuals who need the most academic support, thereby diverting failure or at the very least, reducing the chance of this occurring. Intervention could include mentoring, tutoring and specially designed pedagogical programmes.

Improved assessment design: Teachers, researchers and lecturers should leverage the power of data-driven techniques in setting standards and assessment refinement. Instruments with more predictive power will also be better indicators of student competencies and needs.

Individualised learning pathways: When instruments are well developed, possess good predictive power and are aligned with curriculum goals, it is possible to tailor instruction or interventions to specific and individual needs.

Scientific, evidence-based decision-making: Whenever high-stakes decisions are made, researchers and educators should strive to use rigorous, scientifically backed empirical data as the basis for the decision-making process. The current article supports this aim of empowered and sound decision-making by offering analysis techniques, which has been shown works with well-designed instruments.

Summary and the way forward

In the current study, the Rasch model was used to gauge the functioning of the assessments and confirm instrument reliability and validity. The Rasch model was also used to transform the ordinal total score into an interval scale. The current paper shows that school monitoring assessments can identify academically at-risk learners and alert schools and teachers to help those who need it most. As always, such results should be used with the school assessments and teacher experience and knowledge of the learners. Binomial logistic regression was used to identify the instruments with predictive power to set cut scores and predict group membership. Receiver operating curves were also used to gauge the instruments' usefulness and set cut scores. Both methods predicted group membership, that is, having tertiary access or not having access, beyond chance. The results demonstrate that binomial logistic regression and receiver operating curves can be used to set cut scores and predict which learners are academically at risk. The results could be enhanced by using multiple sources of information, such

as school marks, teacher evaluations and peer ratings. When such methods are combined with multiple data sources, more accurate cut scores can be set and mastery levels within the tests can be identified. Using panels is time-consuming and may require other resources not available to the test designer, such as funds, access to field specialists and knowledge of the methods (Engelhard 2013). This chapter examines how to set cut scores when resources are unavailable and shows reasonable alternative methods using statistical and predictive validity.

The current study demonstrates that scientifically setting cut scores through logistic regression and ROC analysis is a viable alternative when traditional, expert-driven methods are impractical. These approaches enhance the objectivity and reliability of educational assessments and have substantial implications for policy, practice and research. Educators and policymakers can make informed decisions that improve student outcomes and promote equitable access to educational opportunities by adopting data-driven methodologies.

Based on the data and findings presented, researchers are encouraged to use the following guidelines when applying the models to their datasets:

- **Try the models in varied educational contexts:** Applying binomial logistic regression and ROC in diverse educational milieu, including low-income schools, will help assess the generalisability of the current manuscript.
- **Integrate binomial logistic regression and ROC with additional information about learners and students**, which could include teacher evaluations and school-related data. That way, a more holistic and accurate measurement can be obtained.
- **There is a need to gauge the longitudinal usefulness of the models;** researchers could conduct studies that track learner progression throughout high school to university to determine how well the predictions hold up.
- **Explore across subject-specific variability:** The current study used data from mathematics and science, and there is a need to assess whether binomial logistic regression and ROC would work just as well in the soft sciences.

- **The ethical implications** should be explored when a data-driven cut-score setting is used, especially when any high-stakes decisions are made or interventions are made available only for specific groups. Other issues that require further investigation include data privacy, bias mitigation and student autonomy in educational decision-making.

Limitations

The demonstrated effectiveness of ROC relies on having more than one accurate predictor. Statistical knowledge would also be necessary to run the analysis, and this may not always be accessible to educators. Well-designed, refined and piloted instruments may also not be available to educators, or they may not have sufficient knowledge and time to design such assessments. While the current paper opts to offer an alternative to labour-intensive cut-score setting, there is also an assumption of cultural and intellectual capital, which may not be available in developing settings.

References

Angoff, W. H. 1971. Scales, norms, and equivalent scores. In: *Educational Measurement,* edited by R. L. Thorndike. Washington, DC: American Council on Education.

Arce, A. J. and Wang, Z. 2012. Applying Rasch model and generalizability theory to study modified-Angoff cut scores. *International Journal of Testing,* 12(1): 44–60.

Baghaei, P. 2007. Applying the Rasch rating-scale model to set multiple cut-offs. *Rasch Measurement Transactions*, 20(4): 1075–1076. [Online] Available from: *http://www.rasch.org/rmt/rmt204a.htm* [Accessed on 12 March 2025].

Baghaei, P. 2009. A Rasch-informed standard setting procedure. *Rasch Measurement Transactions*, 23(2): 1214. [Online] Available : *http://www.rasch.org/rmt/rmt232f.htm* [Accessed on 12 March 2025].

Biddle, R. E. 1993. How to set cutoff scores for knowledge tests used in promotion, training, certification, and licensing. *Public Personnel*

Management, 22(1).

Blömeke, S. and Gustafsson, J.-E. 2017. *Standard setting in education: The Nordic countries in an international perspective.*

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. and Krathwohl, D. R. 1956. *Handbook I: Cognitive domain.* New York: David McKay Company.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R. and Young, S. L. 2018. Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health,* 6.

Bond, T. G., Yan, Z. and Heene, M. 2020. *Applying the Rasch model: Fundamental measurement in the human sciences.* 4th edition. New York, NY: Routledge. [Online]. Available at: *https://www.taylorfrancis.com/books/9780429030499* [Accessed on 23 March 2025].

Boone, W. J. 2016. Rasch analysis for instrument development: Why, when, and how? *CBE-Life Sciences Education,* 15(4): rm4.

Boone, W. J., Staver, J. R. and Yale, M. S. 2014. *Rasch analysis in the human sciences.* New York: Springer.

Bowers, J. J. and Shindoll, R. R. 1989. *A comparison of the Angoff, Beuk, and Hofstee methods for setting a passing score.* American College Testing Program.

Camasso, M. J. and Jagannathan, R. 1995. Prediction accuracy of the Washington and Illinois risk assessment instruments: An application of receiver operating characteristic curve analysis. *Social Work Research,* 19(3): 174–183.

Carlson, J., Tomkowiak, J. and Stilp, C. 2009. Using the Angoff method to set defensible cutoff scores for standardized patient performance evaluations in PA education. *The Journal of Physician Assistant Education,* 20(1): 15–23.

Cizek, G. J., Bunch, M. B. and Koons, H. 2004. *Setting performance standards: Contemporary methods.* NCME. [Online]. Available at: *http://ncme.org/linkservid/8188D217-1320-5CAE-6EA9C0FC1232764F/showMeta/0/* [Accessed on 12 March 2025].

Cohen-Schotanus, J. and Van der Vleuten, C. P. 2010. A standard setting method with the best performing students as point of reference: practical and affordable. *Medical Teacher,* 32(2): 154–160.

Combrinck, C. 2018. *The use of Rasch measurement theory to address measurement and analysis challenges in social science research.* PhD thesis, University of Pretoria, South Africa. [Online]. Available at: *http://hdl.handle.net/2263/67982* [Accessed on 13 February 2025].

Combrinck, C. 2018. 2020. Is this a useful instrument? An introduction to Rasch models for evaluating tests and questionnaires. In: (eds.). *Online Readings in Research Methods (ORIM),* eidted by S. Kramer, S. Laher, A. Fynn, and H. H. Janse Van Vuuren. Pretoria, South Africa: Psychological Society of South Africa. pp. 127–181. [Online]. Available at: *https://www.psyssa.com/newsroom/publications/orim/chapter-6-is-this-a-useful-instrument/* [Accessed on 17 October 2023].

Combrinck, C., Scherman, V. and Maree, D. J. 2016. The use of Rasch competency bands for reporting criterion-referenced feedback and curriculum-standards attainment. *Perspectives in Education,* 34(4): 62–78.

Combrinck, C., Scherman, V. and Maree, D. J. 2017. Evaluating anchor items and reframing assessment results through a practical application of the Rasch measurement model. *South African Journal of Psychology,* 47(3): 316–329.

Combrinck, C., Scherman, V., Maree, D. and Howie, S. 2018. Multiple imputation for dichotomous MNAR items using recursive structural equation modeling with Rasch measures as predictors. *SAGE Open,* 8(1): 1–12.

Costello, G. R., Davis, K. R. and Crocco, O. S. 2022. Learning by doing: Student & faculty reflections on a collaborative model for conducting and publishing mixed methods research in a graduate course. *Innovative Higher Education,* 47(6): 1067–1084.

Deno, S. L. 2003. Developments in curriculum-based measurement. *The Journal of Special Education,* 37(3): 184 - 192.

Department of Basic Education (DBE). 2024. *Curriculum Assessment Policy Statements (CAPS).* [Online]. Available at: *https://www.education.gov.za/Curriculum/CurriculumAssessmentPolicyStatements(CAPS).aspx* [Accessed on 02 June 2025].

Diniz, M. A. 2022. Statistical methods for validation of predictive models. *Journal of Nuclear Cardiology,* 29(6): 3248–3255.

DiStefano, C. and Morgan, G. 2011. Examining classification criteria: A

comparison of three cut score methods. *Psychological Assessment,* 23(2): 354–363.

Engelhard Jr, G. 2013. *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* Routledge.

Field, A. P. 2009. *Discovering statistics using SPSS.* London: Sage.

— — —. 2024. *Discovering statistics using IBM SPSS statistics.* Sage publications limited.

Gervais, J. 2016. The operational definition of competency-based education. *The Journal of Competency-Based Education,* 1(2): 98–106.

Grosse, M. E. and Wright, B. D. 1987. Criterion item standard setting. *Institute for Objective Measurement.* [Online]. Available at: *http://www.rasch.org/memo84.pdf* [Accessed on 11 June 2016].

Habibzadeh, F., Habibzadeh, P. and Yadollahie, M. 2016. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia Medica,* 26(3): 297–307.

Hintze, J. M. and Silberglitt, B. 2005. A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review,* 34(3): 372–386.

Hosmer, D. W., Hosmer, T., Le Cessie, S. and Lemeshow, S. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics In Medicine,* 16(9): 965–980.

IBM. 2025. *IBM SPSS Statistics for Windows (Version 30.0).* [Online]. Available at: *https://www.ibm.com/spss* [Accessed on 09 August 2024].

Kaftandjieva, F. 2010. *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL.* Ealta, Cito, Arnhem.

Kellow, J. T. and Wilson, V. L. 2008. Setting standards and establishing cut scores on criterion-referenced assessments: Some technical and practical considerations. In: *Best Practice in Quantitative Methods,* edited by J. W. Osborne. United States of America: Sage. pp. 151–160.

Khalid, M. N., Shafiq, F. and Ahmed, S. 2022. A comparison of standard setting methods for setting cut-scores for assessments with constructed response questions. *Pakistan Journal of Educational Research and Evaluation (PJERE),* 9(2).

Khatimin, N., Aziz, A. A., Zaharim, A. and Yasin, S. H. M. 2013.

Development of objective standard setting using Rasch measurement model in Malaysian institution of higher learning. *International Education Studies,* 6(6): 151–160.

Kline, P. 2015. *A handbook of test construction (psychology revivals): Introduction to psychometric design.* Routledge.

Klingbeil, D. A., McComas, J. J, Burns, M. K. and Helman, L. 2015. Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools,* 52(5): 500–514.

Kohrt, B. A. and Kaiser, B. N. 2021. Measuring mental health in humanitarian crises: A practitioner's guide to validity. *Conflict and Health,* 15(1): 72.

Kubiszyn, T. and Borich, G. D. 2024. *Educational testing and measurement.* John Wiley & Sons.

Laher, S. and Cockcroft, K. 2017. Moving from culturally biased to culturally responsive assessment practices in low-resource, multicultural settings. *Professional Psychology: Research and Practice,* 48(2): 115.

Latif, E. and Zhai, X. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence,* 6: 100210.

Linacre, J. M. 1994. Evaluating a ROC screening test. *Rasch Measurement Transactions,* 7(4): 317–318. [Online]. Available at: *http://www.rasch.org/rmt/rmt74a.htm* [Accessed on 27 May 2025].

Linacre, J. M. 2023a. *Winsteps® (Version 5.4.0.0).* [Online]. Available at: *http://0-dx.doi.org.innopac.up.ac.za/https://www.winsteps.com/* [Accessed on 27 May 2025].

Linacre, J. M. 2023b. *Winsteps® Rasch measurement computer program User's Guide.* Portland, Oregon: Winsteps.com. [Online]. Available at: *https://www.winsteps.com/* [Accessed on 27 May 2025].

Longford, T. N. 1996. Reconciling experts' differences in setting cut scores for pass-fail decisions. *Journal of Educational and Behavioral Statistics,* 21(3): 203–213.

MacCann, R. G. and Stanley, G. 2019. The use of Rasch modeling to improve standard setting. *Practical Assessment, Research, and Evaluation,* 11(1): 2.

Masters, G. N. 2016. Partial credit model. In: *Handbook of item response theory,* edited by W. J. van Der Linden. Chapman and Hall/CRC. pp.

109–126.

Metz, C. E. 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine,* 8(4): 283–298.

Mitchell, M. 2019. *Academic screening in middle school: How well do AIMSweb measures of oral reading fluency, and NWEA measures of academic progress, predict future performance on state exams.* PhD thesis, University of Wisconsin, USA.

Myers, R. 1990. *Classical and modern regression with applications.* 2nd ed. Boston: Duxbury.

Park, S. Y., Lee, S.-H., Kim, M.-J., Ji, K.-H. and Ryu, J. H. 2021. Comparing the cut score for the borderline group method and borderline regression method with norm-referenced standard setting in an objective structured clinical examination in medical school in Korea. *Journal of Educational Evaluation for Health Professions,* 18: 25.

Parsaeian, M., Khodaie, E., Izanloo, B. and Salehi, K. 2024. Determining the cut-off score of criterion-referenced tests using non-parametric estimation methods of the Youden index. *Educational Measurement,* 14(56).

Pitoniak, M. J. and Morgan, D. L. 2017. Setting and validating cut scores for tests. In: *Handbook on measurement, assessment, and evaluation in higher education*: Routledge. pp. 235–258.

Ricker, K. L. 2006. Setting cut-scores: A critical review of the Angoff and modified Angoff methods. *The Alberta Journal of Educational Research,* 52(1): 53–64.

Schultz, E. M. 2006. Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practice,* 25(3): 4–13.

Silberglitt, B. and Hintze, J. M. 2005. Formative assessment using oral reading fluency cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment,* 23(4): 304–325.

Smith, E. V. and Stone, G. E. 2009. *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement modes.* United States of America: JAM Press Books.

Spaull, N. 2015. Accountability and capacity in South African education.

Education as Change, 19(3): 113–142.
Stone, G. 2011. Standard Setting, Cut-Scores, and Incorrect Decisions. *Rasch Measurement Transactions*, 24(4):1311. [Online]. Available at: *http://www.rasch.org/rmt/rmt244g.htm* [Accessed on 13 May 2016].
Stone, G., Koskey, K. L. K. and Sondergeld, T. A. 2011. Comparing construct definition in the Angoff and objective standard setting models: Playing in a house of cards without a full deck. *Educational and Psychological Measurement: Interdisciplinary Research and Perspectives.* [Online]. Available at: *https://doi.org/10.1177/0013164410394338* [Accessed on 24 April 2016].
Tabachnick, B. G. and Fidell, L. S. 2007. *Using multivariate statistics.* United States of America: Pearson Education.
Tannenbaum, R. J. and Kannan, P. 2015. Consistency of Angoff-based standard-setting judgments: are item judgments and passing scores replicable across different panels of experts? *Educational Assessment,* 20(1): 66–78.
Umalusi. 2014. *Certification and verification.* [Online]. Available from: *https://www.umalusi.org.za/services/verification/* [Accessed on 1 May].
Wang, D. and Keller, L. A. 2024. Using ROC analysis to refine cut scores following a standard setting process. *Educational and Psychological Measurement*: 85(2): 313–335.
Wang, N. 2003. Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement,* 40(3): 231–253.
Wright, B. D. 2000. How to set standards. *Rasch Measurement Transactions,* 14(1): 740. [Online]. Available at: *http://www.rasch.org/rmt/rmt141n.htm* [Accessed on 11 September 2024].
Wright, B. D. and Grosse, M. 1993. *How to set standards.* [Online]. Available from: *http://www.rasch.org/rmt/rmt73e.htm* [Accessed on 31 August 2025].
Wright, B. D. and Stone, M. H. 1979. *Best test design.* Chicago: MESA Press.
Wyse, A. E. 2017. Five methods for estimating Angoff cut scores with IRT. *Educational Measurement: Issues and Practice,* 36(4): 16–27.
— — —. 2018. Examining how professional roles and test development experiences impact Angoff ratings. *Applied Measurement in Education,* 31(4): 324–334.

Yousef, M., Alshawwa, L., Tekian, A. and Park, Y. 2017. Challenging the arbitrary cutoff score of 60%: Standard setting evidence from preclinical operative dentistry course. *Medical Teacher,* 39(sup1): S75–S79.

Yudkowsky, R., Park, Y. S., Lineberry, M., Knox, A. and Ritter, E. M. 2015. Setting mastery learning standards. *Academic Medicine,* 90(11): 1495–1500.

Zweig, M. H. and Campbell, G. 1993. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry,* 39(4): 561-577.