Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes[1,2]

Alex J. Bowers
Teachers College, Columbia University
Bowers@tc.edu

Xiaoliang Zhou
Teachers College, Columbia University
xz2256@tc.columbia.edu

ABSTRACT:

Early Warning Systems (EWS) and Early Warning Indictors (EWI) have recently emerged as an attractive domain for states and school districts interested in predicting student outcomes using data that schools already collect with the intention to better time and tailor interventions. However, current diagnostic measures used across the domain do not consider the dual issues of sensitivity and specificity of predictors, key components for considering accuracy. We apply signal detection theory using Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) analysis adapted from the engineering and medical domains, and using the pROC package in R. Using nationally generalizable data from the Education Longitudinal Study of 2002 (ELS:2002) we provide examples of applying ROC accuracy analysis to a variety of predictors of student outcomes, such as dropping out of high school, college enrollment, and postsecondary STEM degrees and careers.

Keywords: ROC, AUC, Early Warning System, Early Warning Indicator, signal detection theory, dropout, college enrollment, Postsecondary STEM Degree, hard STEM career, soft STEM career

BACKGROUND and LITERATURE REVIEW:

Using Early Warning Systems (EWS) and Early Warning Indicators (EWI) to predict student outcomes has become an emerging domain of interest (Agasiti & Bowers, 2017; Allensworth, 2006; Allensworth, & Easton, 2007; Allensworth, Nagaoka, & Johnson, 2018; Allensworth, Gwynne, Moore, & de la Torre, 2014; Bowers, Sprott, & Taff, 2013; Faria et al., 2017; Kemple, Segeritz, & Stephenson, 2013; Ocumpaugh et al., 2017). EWS and EWI are defined as "an intentional process whereby school personnel collectively analyze student data to monitor students at risk of falling off track for graduation and to provide the interventions and resources to intervene" (Davis, Herzog, & Legters, 2013, p. 84). EWS and EWI have been applied to predicting education outcomes, such as high school dropout (Knowles, 2015; Stuit et al., 2016; Tamhane, Ikbal, Sengupta, Duggirala, & Appleton, 2014), successful middle school to high school transition (Allensworth et al., 2014; Faria et al., 2017), college readiness (Hughes, & Petscher, 2016; Koon, & Petscher, 2016; Phillips et al., 2015), and academic performance (Ikbal et al., 2015; Lacefield & Applegate, 2018; Lacefield, Applegate, Zeller, & Carpenter, 2012; Macfadyen, & Dawson, 2010) to name but a few. EWS and EWI are becoming increasingly popular because they can identify a small set of variables that allow educators to make timely decisions (Frazelle, & Nagel, 2015; Macfadyen, & Dawson, 2010), that can inform personalized interventions and help to reduce student retention and dropout (Ikbal et al., 2015; Tamhane, Ikbal, Sengupta, Duggirala, & Appleton, 2014), and that "often prompt[s] improvements in a district's system of supports" (Supovitz, Foley, & Mishook, 2012, p. 2). For example, randomly assigning 73 schools to treatment and control groups to use an Early Warning Intervention and Monitoring System (EWIMS), Faria et al. (2017) found that EWIMS was helpful for improving course completion and student attendance, each of which are major indicators of students being off track for graduation (Allensworth & Luppescu, 2018).

Soland (2013) examined the accuracy of an EWS with the nationally representative National Education Longitudinal Survey (NELS) data of high school students, teachers, and parents from 1988-2000, predicting dropout and college going. After doing several longitudinal logistic analyses, Soland (2013) found that

> …EWS forecasts could be valuable both as organizational tools and for their precision. For example, model predictions of college going were more accurate than teacher predictions and remained accurate for students about whom teachers were unsure or disagreed. (p.259)

Soland's (2013) study is a recent and robust investigation into the performance of an EWS. However, as with the vast majority of the literature in this domain, missing from this conversation is an examination of the accuracy of the predictors used in the early warning system. More recently, in a follow-up

[1] This document is a pre-print of this manuscript, published in the *Journal of Education for Students Placed At Risk* (JESPAR). Citation: Bowers, A.J., Zhou, X. (2019) Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. *Journal of Education for Students Placed At Risk,* 24(1) 20-46. https://doi.org/10.1080/10824669.2018.1523734

[2] This research was supported by a grant from the National Science Foundation (NSF IIS-1546653). Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of funding agencies.

study Soland (2017) used a decision tree and a lasso regression analysis framework with *n*=9,601 students from the NELS sample from 1988-2000. In this study, the author worked to reduce thousands of measures with cross-validation techniques down to a set of variables measuring SES, postsecondary aspirations, academic readiness, and teachers' perceptions of readiness. The author then used the reduced set of variables to predict students enrolling in college and persisting for a semester, noting that his model achieved almost 90% accuracy. However, Soland (2017) calculated classification accuracy with two separate measures, one dividing the number of correct predictions by the total number of students, the other dividing false positives by false negatives. Although the second method used false positives and false negatives, the two metrics do not take into account the current literature on accuracy of predictors from the research domain of signal detection theory. The core issues of signal detection theory in assessing predictor accuracy is a central missing detail across almost all of the current research in the EWI/EWS domain.

Signal detection theory is a domain of research which originated in engineering and medicine (Gönen, 2007; Hanley & McNeil, 1982; Swets, 1988; Swets, Dawes, & Monahan, 2000; Zwieg & Campbell, 1993) and has more recently been applied to domains such as law, education, and youth development to understand the accuracy of at-risk indicators and predictors (Bowers, Sprott & Taff, 2013; Knowles, 2015; Olver, Stockdale, & Wormith, 2009; Rice & Harris, 2005; Vivo & Franco, 2008). A core measure of accuracy within signal detection theory is the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), in which the sensitivity of a predictor (the true-positive proportion) is compared to the specificity (the true-negative proportion), a calculation which is the central concern of the present study and which we review and provide example applications in more detail below. Although there have been some reviews of metrics for evaluating the accuracy of student at-risk prediction models (National Center on Response to Intervention, 2010; Pelánek, 2014; Pelánek, 2015) or case studies on the use of Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) in human computer interaction research (Fogarty, Baker, & Hudson, 2005), only a few studies have used ROC to examine the accuracy of diagnostic measures for predictors of outcomes in education research (e.g., Carlson, 2018; D'Agostino, Rodgers, & Mauck, 2018; Johnson & Semmelroth, 2010; Jordan, Glutting, Ramineni, & Watkins, 2010; Liao, Yao, Chien, Cheng, & Hsieh, 2014; Nicholls, Wolfe, Besterfield-Sacre, & Shuman, 2010; Stuit et al., 2016). Research studies in education have rarely tested the accuracy of flags that are used to predict students at risk of specific outcomes (Bowers, Sprott, & Taff, 2013), which has resulted in a lack of reported accuracy for predictors across this domain.

Thus, the question for EWI and EWS in education is how to identify and report accurate predictors of students at-risk. Education researchers, policymakers, and practitioners need a means to measure the accuracy of diagnostic systems, as misidentification of students at risk of negative outcomes has two broad drawbacks. As Gleason and Dynarski (2002) noted in reference to this issue of accuracy with early warning indicators of dropping out of high school, low accuracy EWIs lead to students who are identified as at risk of dropping out but who would have never dropped out. These students receive dropout interventions, but as noted by the authors, if the students would have not dropped out, perhaps these resources could be spent in better ways, in addition to the problem of negatively labeling a student as at risk with an inaccurate predictor. Conversely, students who will drop out but are never identified by the indicator do not receive needed resources and interventions that could help them persist in school. Importantly for this domain, in reviewing 110 dropout flags from the literature, Bowers et al. (2013), showed that the vast majority of EWIs in the dropout domain were not much better than a 50-50 guess, thus misidentifying large percentages of students as dropping out when they would not have dropped out, as well as missing a large percentage of students who will drop out but are never identified. The authors term this second set of students the "quiet dropouts" (Bowers & Sprott, 2012a, 2012b).

In the present study, we detail the use of a measure of accuracy known as the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), drawn from Signal Detection Theory in the Engineering and Medical domains. Whereas ROC AUC is commonly used to evaluate the accuracy of predictors in fields such as medicine, radiology, engineering, data mining, and machine learning, it is a relatively new concept for education early warning systems.

Signal Detection Theory

Signal detection theory deals with sensitivity and specificity. In Figure 1 we adapt a contingency table from the previous literature, noting how to compute sensitivity, specificity, and other related diagnostics (Swets, 1988, p. 1286; Bowers, Sprott, & Taff, 2013, p. 83

The upper part of the figure is a contingency table, with rows as predictions and columns as events. The table also shows the row and column marginal sums and total sum. The lower part of the figure is the calculation of measures of interest for accuracy. Discussing the contingency table in Figure 1 using Type I error and Type II error can help to understand the calculations for sensitivity and specificity. Here *Sensitivity* is analogous to "1-Type II error" is defined as the probability of identifying positive events as positive (True-positive Proportion), so 1-*Sensitivity* is analogous to "Type II error" as the probability of identifying positive events as negative (False-negative Proportion). *Specificity* analogous to "1-Type I error" is defined as the probability of identifying negative events as negative (True-negative Proportion), so 1-*Specificity* analogous to "Type I error" is the probability of identifying negative events as positive (False-positive Proportion). Another related concept *precision* (Positive Predictive Value) is defined as the probability of correctly predicting positive events. In other words for example, for an EWI predicting dropping out of high school, sensitivity can be thought of as the "hits" where it is the proportion of students predicted to drop out who actually did

		Observation	
		Positive	Negative
Prediction	Positive	True-positive (*a*)	False-positive (*b*)
	Negative	False-negative (*c*)	True-negative (*d*)

Accuracy	$(a + d) \ / N$
Precision	$a / (a + b)$
Sensitivity (Recall / True-positive Proportion)	$a / (a + c)$
Specificity (True-negative Proportion)	$d / (b + d)$
1- Specificity (False-positive Proportion)	$b / (b + d)$
Kappa	$(\text{Accuracy} - \text{R}) / (1 - \text{R})$
	$\text{R} = ((a + c)(a + b) + (b + d)(c + d)) / N^2$

Figure 1: *Confusion Table for Calculating Contingency Proportions*

drop out, while 1-specificity can be thought of as the "false alarms" which is the proportion of students who had the dropout predictor but still graduated.

As noted in the signal detection theory literature (Gönen, 2007; Hanley & McNeil, 1982; Swets, 1988; Swets et al., 2000; Zwieg & Campbell, 1993), the central issue in the accuracy of predictors of outcomes is that both the sensitivity and specificity must be taken into account, as these two dimensions of the true-positive proportion (sensitivity) and the false-postive proportion (1-specificity) act in concert when discussing accuracy (Bowers et al., 2013). Said another way, for an early warning indicator to be accurate, the net that is cast should catch the intended targets (it must be sensitive) while not misidentifying the wrong targets (1-specificity, the opposite of specificity). Interestingly, these two dimensions of accuracy can then be plotted in an x,y coordinate plane, with the sensitivity and 1-specificity for each early warning indicator plotted in these two dimensions. In this way, a Receiver Operating Characteristic (ROC) plot incorporates sensitivity and specificity into one figure, making it possible for researchers to visually compare the accuracy of different indicators (Allensworth et al, 2014; Bowers et al., 2013; Knowles, 2015; Swets, 1988; Swets et al., 2000). The false-positive proportion (1-specificity) is on the x-axis and the true-positive proportion (sensitivity) is on the y-axis. For any predictor, a change in specificity leads to a corresponding change in sensitivity, and this relation is plotted as a curve in the ROC plot. The Area Under the Curve is called ROC AUC, and ranges from 0 to 1.0, with 0.5 as a 50-50 guess and 1.0 as a predictor that is 100% accurate. As a diagnostic measure, the AUC calculation takes into account both the true-positive proportion and the false-positive proportion, and can be compared to identify which predictors are more accurate than others, and which are closest to a 50-50 guess, or worse. An accurate predictor should have a large AUC (closer to point 0,1 on the ROC AUC plot with an AUC above 0.5 and closer to 1.0), which means that the predictor performs well in both dimensions of sensitivity and specificity. While the broader ROC AUC literature encourages researchers and practitioners to compare AUCs of individual diagnostic predictors within the same ROC analysis to identify the most accurate predictors (Bowers et al, 2013; Hanley & McNeil, 1982; Swets, 1988; Swets et al., 2000; Zwieg & Campbell, 1993) certain fields have established rules of thumb for levels of accuracy as determined by AUC. For example, in screener validity analysis in the Response to Intervention field (RTI) (D'Agostino, Rodgers, & Mauck, 2018), the National Center on Response to Intervention (2010) (NCRTI) as their "Technical Standard 1" has indicated that for classification accuracy of screening tests for RTI, that an AUC above 0.85 is considered "convincing evidence" of classification accuracy, between 0.75 and 0.85 is "partially convincing evidence" and less than 0.75 is "unconvincing evidence".

Framework and Research Questions

Currently, there are only a small number of studies in education using ROC or AUC as a diagnostic measure for evaluating predictors in the EWI/EWS literature. Even for those studies that do use ROC (Allensworth et al., 2014; Torres, Bancroft, & Stroub, 2015) or AUC (Becker et al., 2014; Cummings, & Smolkowski, 2015; Laracy, Hojnoski, & Dever, 2016; Knowles, 2015; Vivo, & Franco, 2008; Wilson et al., 2016) to evaluate predictors, researchers used the technique among several other measures and did not necessarily consider ROC as more accurate than other measures in terms of measuring a predictor's accuracy. One exception is the research by Horn and Lee (2017) who advocated the use of sensitivity, specificity, and precision together to judge the accuracy of their performance indicators. Bowers, Sprott and Taff (2013) provide a comprehensive study aimed to demonstrate the accuracy of ROC as a diagnostic measure in which the authors analyzed 110 high school dropout flags from the dropout prediction literature using a ROC analysis. They illustrated the possibility of visually comparing predictor accuracy in terms of both sensitivity and specificity. The Bowers et al. (2013) study demonstrated how researchers and practitioners could use a ROC plot to graphically display and compare the accuracy of dichotomous flags as a point in the ROC space (such as high suspensions, failed English, failed mathematics, high absences).

For example, the study concluded that the most accurate cross-sectional single time point dropout flag in the literature was the Chicago on-track indicator which includes low course credits and failing at least one core course in ninth grade (Allensworth, 2013; Allensworth & Easton, 2007), while more accurate predictors each used longitudinal predictors and growth mixture modeling to identify significantly different trajectories of student performance (Bowers et al, 2013), such as declining trajectories of non-cumulative grade point average in the first three semesters of high school (Bowers, 2019; Bowers & Sprott, 2012b; Brookhart et al., 2016). However, a critical missing component for use of these measures in EWS was that Bowers et al. (2013) did not consider the use of AUC in the ROC accuracy analysis. The signal detection theory literature (Swets, 1988) demonstrates that indicator accuracy can be better compared through plotting continuous variables (rather than just dichotomous indicators) within the ROC space and then assessing the area under the curve (AUC) for each continuous variable in which a larger AUC indicates a higher level of accuracy. Additionally, AUC provides a means to statistically compare the AUC of two different indicators to assess the extent to which they are statistically significantly different (DeLong, DeLong, & Clarke-Pearson, 1988; Hanley & McNeil, 1983).

With increasing attention in the education analytics literature to creating and analyzing early warning indicators and systems (Agasisti & Bowers, 2017; Bowers, 2017; Piety, Hickey, & Bishop, 2014), there is a need to provide a discussion of how to assess and compare accuracy of indicators in the EWI/EWS domain using ROC AUC with examples and code to help inform current and future research, policy and practice. In this study we select a range of early warning predictors and outcomes of interest using publically available nationally-generalizable data from the U.S Department of Education National Center for Education Statistics and drawing on the EWI/EWS literature noted above, and provide the ROC AUC analysis for each using open source software, comparing multiple early warning indicators for dropping out of high school, post-secondary college enrollment, and high school students eventually choosing a career in STEM (science, technology, engineering, or mathematics) by age 26.

METHODS:

Data

This study is a secondary analysis of the publically accessible Education Longitudinal Study of 2002 (ELS:2002), a US high school student sample collected by the National Center for Education Statistics (NCES) from 2002 to 2012 (Ingles et al., 2014). ELS:2002 includes a nationally generalizable sample of about 16,000 students who were in grade 10 in 2002 across 750 high schools. For the present analysis the sample sizes of our selected variable range from 10,511 (college enrollment) to 16,197 (dropout, one or more flags). We selected ELS:2002 as it is currently the most recent comprehensive ten-year longitudinal U.S. high school student survey of its type. Appendix A lists the descriptive statistics and labels of the variables analyzed in this study.

Variables included in the Analysis: Predictors of Dropout, College Enrollment, Postsecondary STEM Degree, and Hard/Soft STEM Occupation

We drew on the current EWI/EWS literature to inform our variable selection for this study, including the outcomes of interest of 1) dropping out of high school, 2) college enrollment, and 3) postsecondary STEM degree, 4) STEM career occupation at age 26, as well as the indicators used to predict these outcomes. We selected to focus on both secondary and post-secondary schooling indicators and outcomes for our examples of ROC AUC as the EWI/EWS literature is developing rapidly for K-12 and colleges and universities in the education analytics, learning analytics, education data science, and academic analytics domains, especially in relation to student STEM outcomes, as researchers and practitioners wish to identify accurate indicators of positive transitions from high school, through college, into careers (Agasisti & Bowers, 2017; Baker, 2013; Bowers, 2017; Knight & Shum, 2018; Krumm, Means, & Bienkowski, 2018; Piety, Hickey, & Bishop, 2014; Siemens, 2013). Nevertheless, as noted above, across this literature there are few examples of providing accuracy analysis, however as the literature across these domains uses multiple predictors for early warning indicators in each of the four outcomes of interest in this study, we draw on this literature to inform our variable selection.

The first education outcome we examine is high school dropout. Following Bowers et al. (2013), our first set of indicators to predict high school dropout mirrors the dichotomous indicators from Balfanz, Herzog, and Mac Iver (2007) of students in sixth grade, including attendance, suspension, misbehavior, failed math, failed English, any one or more of these flags, any one flag, any two flags, any three flags, and all four flags. The Balfanz et al. (2007) study is an important study to start with as a focus, as it examines a large sample of sixth graders from Philadelphia, is widely cited, and these dropout flags are used in many state policy recommendations for at-risk prediction systems (Rumberger et al., 2017). However, currently there are no public national datasets that include the sixth grade as well as overall high school and career outcomes, thus we adapted the Balfanz et al. (2007) dropout flags to the ELS:2002 dataset. As we selected the ELS:2002 dataset for this study, ELS:2002 includes a representative sample of students in grade 10 from across 750 high schools in 2002. To provide an example of ROC AUC accuracy analysis with similar variables to Balfanz, Herzog, and Mac Iver (2007) using the public ELS:2002 data, we included the variables "1st quantile standardized math score" where we dichotomized students' standardized math scores with the cutoff point being the 1st quantile score. To follow Balfanz et al. (2007), we dichotomized both "absent" and "suspension" to derive composite flags with Boolean operators such as *and* and *or*. In addition to the Balfanz et al. (2007) predictors of dropout, we examined the accuracy of multiple continuous predictors of dropout from the literature, such as attendance (Adelman, 2006; Allensworth, Gwynne, Moore, & de la Torre, 2014; Allensworth, Nagaoka, & Johnson, 2018;

Balfanz & Boccanfuso, 2007; Bowers, & Sprott, 2012a, 2012b; Geiser & Santelices, 2007; Kemple et al., 2013; Neild, Balfanz, & Herzog, 2007; Noble & Sawyer, 2004; Quadri & Kalyankar, 2010; Soland, 2013; Soland, 2017), suspension (Balfanz, Herzog, & Mac Iver, 2007; Bowers & Sprott, 2012a; Soland, 2013), extracurricular activities (Mahoney, & Cairns, 1997; Renzulli, & Park, 2000; Soland, 2013), and standardized test scores (Bowers, 2010; Kemple et al., 2013; Soland, 2013; Soland, 2017).

As for accurate predictors for college enrollment, researchers find that high school grades (Allensworth et al., 2018; Bowen, Chingos, & McPherson, 2009; Geiser, & Santelices, 2007; Kemple et al., 2013; Soland, 2013; Soland, 2017), attending at least one Advanced Placement (AP) program (Becker et al., 2014; Soland, 2013), and extracurricular activities (Soland, 2013) are significant predictors for four-year college enrollment. As shown in Appendix A, the sample size for the outcome variable is 10,511. For the outcome variables, "Item legitimate skip/NA" and "Survey component legitimate skip/NA" were considered as in the comparison group, coded as 0 because students having these two options did not receive college education and should be included in the comparison group. pROC, the analysis package used below, carries out listwise deletion on the missing values of both the outcome variable and the predictor before calculating AUC.

For predictors of postsecondary STEM degree (Science Technology Engineering Mathematics), researchers have shown that STEM course selection is a strong predictor of if a student graduates from a postsecondary institution with a STEM degree (Dutta-Moscato, Gopalakrishnan, Lotze, & Becich, 2014). Specifically, we examine the accuracy of the number of STEM courses (Chen, 2013). We also evaluate the accuracy of standardized math score in high school (Chen, 2013; Crisp, Nora, & Taggart, 2009) and STEM course GPA (Chen, 2013; Crisp et al., 2009) in predicting if a student graduates with a postsecondary STEM degree. As shown in Appendix A, the sample size for the outcome variable is 6,936. For the outcome variables, "Item legitimate skip/NA" and "Survey component legitimate skip/NA" were coded as missing data because students having these two options did not receive postsecondary education and should not be included in the comparison group.

We are mixing college and high school variables in predicting occupations at age 26 because previous research literature (Perna et al., 2009) has shown that variables at both high school and college levels are significant predictors of STEM careers. We evaluate the accuracy of standardized math score at grade 10 (Robnett & Leaper, 2013), college STEM course number (Perna et al., 2009), and STEM course GPA in college (Clark Blickenstaff, 2005) in predicting the occupation of a student as in either a "hard STEM" or "soft STEM" career by age 26 at the third follow-up for ELS:2002 in 2012. We draw on the literature in this domain for the definitions of STEM, hard STEM and soft STEM. Here we use the "S.M.A.R.T." definition of STEM courses as courses related to technical fields, foreign languages critical to national security, or qualifying liberal arts (U.S. Department of Education, 2014). We selected this definition for STEM courses in our study as NCES adopted this definition of STEM for ELS:2002. Following the recommendations of the previous literature (Willis, 2013; Whittaker, 2014), typical hard STEM courses include engineering and computer science, whereas typical soft STEM courses include forensic and archaeological science, and the social sciences. As shown in Appendix A, the sample size for the variable of either hard STEM occupations or soft STEM occupations at age 26 is 12,796. For the outcome variables, "Item legitimate skip/NA" and "Survey component legitimate skip/NA" were coded as 0 because students having these two options did not enter STEM careers and should be included in the comparison group.

Receiver Operating Characteristic (ROC) Analysis

The signal detection theory literature (Hanley & McNeil, 1982; Swets, 1988; Swets et al., 2000; Vivo & Franco, 2008; Zwieg & Campbell, 1993) recommends that studies calculate precision, sensitivity, and specificity (see Figure 1). In the contingency table of Figure 1, an event indicates if a student experiences the at-risk outcome (columns), whereas a predictor predicts if the student will have the outcome (rows). According to the signal detection theory literature, precision is defined as the true-positives divided by the number of students predicted to have the flag, and the true-positive proportion (*sensitivity*) as the true-positives divided by the actual number of students at risk. The true-negative proportion (*specificity*) is defined as the true-negatives divided by the number of students without the risk, and the false-positive proportion (1-*specificity*) as the false-positives divided by the number of students without the risk. Indeed, as noted above in this literature (Swets, 1988), two crucial measures of the accuracy of predictors are true-positive proportion ("hits") and false-positive proportion ("false alarms"). Any detection system will always have a trade-off between "hits" and "false alarms", as when one maximizes the number of hits by "casting a wider net", one tends to also increase the number of false alarms (Bowers et al, 2013, p. 83).

Area Under the Curve (AUC)

The focus of the present study is AUC, which has rarely been addressed in the education diagnostics literature. AUC is a step further over ROC analysis, and the purpose of the present study is to encourage the consistent use of AUC for predictors of outcomes in education, with illustrative examples of different variables of risk. AUC is calculated by summing the area under the ROC curve, and the bigger the area, the more accurate the predictor. Formally, the formula for calculating AUC is

$$AUC = \int_0^1 f(x)dx \qquad (1)$$

where $f(x)$ is the function of the ROC curve. However, since $f(x)$ tends not to have an integratable shape like a parabola, methodologists suggest using approximation methods to calculate AUC. For example, Robin et al. (2011) imputed AUC by connecting "empirical ROC points on linear probability

scales" via straight lines and calculating "the subtended area by the trapezoidal rule" (Swets, & Pickett, 1982, p. 31). In other words, Robin et al. divide the x axis to *k* equally spaced intervals, cut the AUC into *k* trapezoids, and add the areas of the trapezoids to approximate the AUC. For feasibility of computation, the approximation approach to calculating AUC or its variants is what is commonly used as the algorithm in ROC AUC software packages.

Significance Testing for AUC Difference

To test if an AUC is significantly different from another one, we need a measure to test AUC difference. One such statistic for significance testing between two continuous predictors is discussed in Hanley and McNeil (1983), with the formula for calculating the *z* score as

$$z = \frac{AUC_1 - AUC_2}{SE^2_{AUC_1} + SE^2_{AUC_2} - 2rSE_{AUC_1}SE_{AUC_2}} \quad (2)$$

where AUC_1 and AUC_2 are the AUC's of the first and second predictors; where SE_{AUC_1} and SE_{AUC_2} are the standard errors of AUC_1 and AUC_2; and where *r* is the correlation between AUC_1 and AUC_2. By comparing the calculated *z* score with corresponding critical *z* value of the standard normal distribution, we will know if the AUC difference of two predictors is significant.

However, the *z* score is suitable only for continuous variables (DeLong et al., 1988), and for non-continuous variables, we need to use other statistics, such as

$$(\widehat{\boldsymbol{AUC}} - \boldsymbol{AUC})\boldsymbol{X}'\left[\boldsymbol{X}\left(\frac{1}{p}\boldsymbol{V}_{10} + \frac{1}{q}\boldsymbol{V}_{01}\right)\boldsymbol{X}'\right]^{-1}\boldsymbol{X}(\widehat{\boldsymbol{AUC}} - \boldsymbol{AUC})' \quad (3)$$

where $\boldsymbol{AUC}$ and $\widehat{\boldsymbol{AUC}}$ are true and estimated AUC vectors; where $\boldsymbol{X}$ is a row vector of coefficients; where *p* and *q* are the numbers of positive and negative cases in reality; and where $\frac{1}{p}\boldsymbol{V}_{10} + \frac{1}{q}\boldsymbol{V}_{01}$ is the estimated covariance matrix for $\widehat{\boldsymbol{AUC}}$. Equation (3) has a chi-square distribution with the degrees of freedom equal to the rank of $\boldsymbol{X}\left(\frac{1}{p}\boldsymbol{V}_{10} + \frac{1}{q}\boldsymbol{V}_{01}\right)\boldsymbol{X}'$. Similar to equation (2), equation (3) has an AUC difference as the numerator and pooled covariance as the denominator. Again, we can compare equation (3) to the corresponding value of the chi-square distribution to test significance of the AUC difference. In the present study, as we discuss below, we use the open source R software with the pROC package (Robin et al, 2011), as pROC adopts this statistic and readers may refer to DeLong et al. (1988) for a more in-depth discussion on the derivation of equation (3).

Comparison with Mann–Whitney–Wilcoxon Test

If approximated by the trapezoid method, then AUC is equivalent to the Mann–Whitney *U*-statistic, or the Mann–Whitney–Wilcoxon, which compares distributions of different samples (Bamber, 1975; DeLong et al., 1988). Since the trapezoid rule underestimates AUC when the number of values of a variable is small (DeLong et al., 1988; Hanley & McNeil, 1983; Swets & Pickett, 1982), DeLong et al. (1988) used the generalized *U*-statistic theory to develop a nonparametric approach to calculating and comparing AUCs with equation (3). Simulation studies (Fanjul-Hevia & González-Manteiga, 2018) show that DeLong et al.'s method produces the best power than other resampling methods when one ROC curve dominates over the other curve. DeLong et al.'s (1988) method is incorporated into pROC (Robin et al., 2011) and the present study selected DeLong et al.'s (1988) approach in the pROC package to compare predictors' AUCs.

AUC vs. Kappa

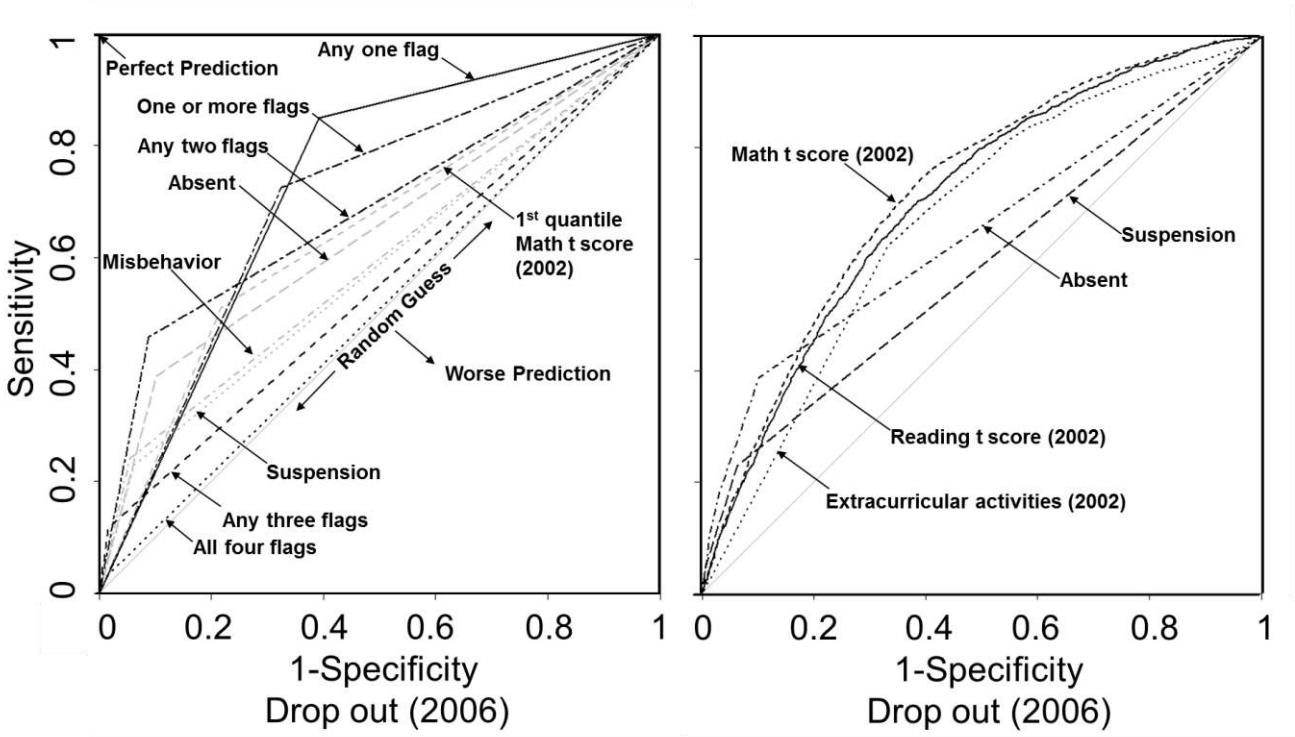
In the education data mining and learning analytics literature, researchers tend to use Kappa rather than ROC AUC (Baker, 2014; Ocumpaugh, Baker, Gowda, Heffernan, & Heffernan, 2014). Kappa measures the degree of agreement between events and predictors. With no agreement, Kappa equals 0 and the relationship between events and predictors is random. With perfect agreement, Kappa equals 1 (Langenbucher, Labouvie, & Morgenstern, 1996).

For the present study, following the recommendations of the signal detection theory literature noted above, we used AUC as the diagnostic measure for evaluating predictors as AUC has advantages over Kappa. First, the meaning of AUC is invariant for different datasets, so a higher value is always better than a lower value (Baker, 2015). Second, AUC provides a more straightforward statistical interpretation (Baker, 2015). Third, AUC can calculate confidence intervals, allowing researchers to compute whether two AUC's are significantly different from each other (Baker, 2015), as we show below in the application examples. Fourth, AUC can compare variables that have some categories dominate over other categories while Kappa cannot handle such variables (Baker, 2015). Fifth, AUC is robust to skewed data, whereas Kappa calculations can be distorted for skewed data (Jeni, Cohn, & De La Torre, 2013).

R Packages

To provide accessible example applications of the use of ROC AUC for education EWI/EWS we rely on the open source statistical R software (R Development Core Team, 2018) and the pROC R package (Robin et al, 2011). Among the several open source R packages for calculation of AUC, we selected the pROC package by Robin et al. (2011) as this R package plots ROC curves, calculates AUC, and tests significance of AUC difference. We provide the R markdown code for the present study in Online Appendix S1 http://doi.org/10.7916/D8K94RDD. pROC allows for significance testing between two predictors and we used OptimalCutpoints (López-Ratón et al., 2014) in R to calculate cutoff points along the ROC curves.

(A) (B)



Predictor:	AUC
Any one flag:	0.729
One or more flags:	0.700
Any two flags:	0.685
1st quantile Math t score (2002):	0.647
Absent:	0.643
Misbehavior :	0.593
Suspension:	0.584
Any three flags:	0.551
All four flags:	0.509

Predictor:	AUC
Math t score (2002):	0.722
Reading t score (2002):	0.706
Extrac. activities (2002):	0.672
Absence:	0.647
Suspension:	0.584

Figure 2: ROC Curves for Predicting Dropout. Panel (A) replicates Balfanz et al. (2007) research on predictors of dropout with ELS:2002 data. The curves in black are for composite flags and those in gray are for raw flags. Panel (B) uses continuous variables to predict dropout with ELS:2002 data.

RESULTS:

The purpose of this study is to provide an overview and introduction of the ROC AUC literature to help inform EWI and EWS research and practice in determining the accuracy of at-risk predictors, and then provide example applications in education using open source software and publically available data. In the below results section, we provide four main examples of ROC AUC using the NCES ELS:2002 dataset and the variables noted in the methods from high school and college to compare the accuracy of at-risk predictors for 1) high school dropout, 2) post-secondary college enrollment, 3) STEM 4-year degree attainment, and 4) hard STEM and soft STEM career outcomes at age 26.

High School Dropout

Figure 2 illustrates the ROC curves for predictors of high school dropout. Panel (A) replicates Balfanz et al.'s (2007) research on predictors of dropout with ELS:2002 data. Some curves, such as misbehavior and suspension, are very close and difficult to distinguish. This is because the predictors are dichotomous, as we wished to attempt to mirror the Balfanz et al. (2007) predictors as closely as we could using the public ELS:2002 dataset to provide a link and entry point from the previous dropout indicator literature to the ROC AUC framework. We can evaluate the accuracy of each predictor through calculating the AUC for each (see methods) and

Table 1: Significance of AUC Difference for Predictors of Continuous Dropout

Predictor 1	Predictor 2	*Z*	*p*-value
Reading t score (2002)	Math t score (2002)	3.220	0.001
Reading t score (2002)	Extracurricular activities (2002)	-3.641	<0.001
Reading t score (2002)	Absent	-33.716	<0.001
Reading t score (2002)	Suspension	-32.105	<0.001
Math t score (2002)	Extracurricular activities (2002)	-5.744	<0.001
Math t score (2002)	Absent	-34.194	<0.001
Math t score (2002)	Suspension	-34.840	<0.001
Extracurricular activities (2002)	Absent	-28.758	<0.001
Extracurricular activities (2002)	Suspension	-28.183	<0.001
Absent	Suspension	5.846	<0.001

comparing the results. As detailed in the signal detection theory literature discussed above (Bowers et al, 2013; Swets, 1988) in a ROC plot such as Figure 2 Panel A, each predictor or indicator is plotted in two dimensions of 1-specificity versus sensitivity. The 45 degree line across the plot indicates a 50-50 random guess, whereas predictors that approach the point 0,1 in the upper left-hand corner approach a more perfect prediction, as point 0,1 indicates a predictor that had no false positives and captured only true positives. For our example here in Figure 2, this would indicate that a predictor perfectly predicted 100% of all students who dropped out, and did not misidentify any students who eventually graduated as dropouts. As noted in the previous literature (Bowers et al, 2013), the vast majority of the predictors in the high school dropout literature fall close to the 45 degree random guess line. For AUC calculations, this is the area under the curve for each indicator and predictor, which ranges from a random guess of 0.5 (half the area of the plot), to 1.0 as a perfect prediction. Predictors with a higher AUC are more accurate. As discussed in the methods, we also provide the statistical pair-wise comparisons for each AUC to determine if each indicator is significantly different from the others in the analysis (see Appendix B).

As shown in Figure 2 Panel A, the composite flags in which dropout flags are combined with Boolean operators were more accurate than individual flags alone by the ROC AUC, as "any one flag" and "one or more flags" each had an AUC of 0.729 and 0.700 respectively. For the other dropout flags these are listed in decreasing order of accuracy with "all four flags" being close to a random guess. These results mirror the previous findings for these flags at grade six (Balfanz et al. 2007; Bowers et al. 2013), and replicate and extend the results to the grade 10 context while calculating the AUC for each predictor for the first time. Appendix B provides the *p*-values for which predictors are significantly different from each other in Figure 2 Panel A. Most pairwise comparisons among the nine predictors are significant, with *p* values less than 0.001. Exceptions are those for "absent" and "1st quantile Math *t* score", "1st quantile Math *t* score" and "any two flags", "misbehavior" and "suspension", and "one or more flags" and "any one flag", as each of these is not significantly different from the other in their accuracy to predict dropping out of high school (see Appendix B).

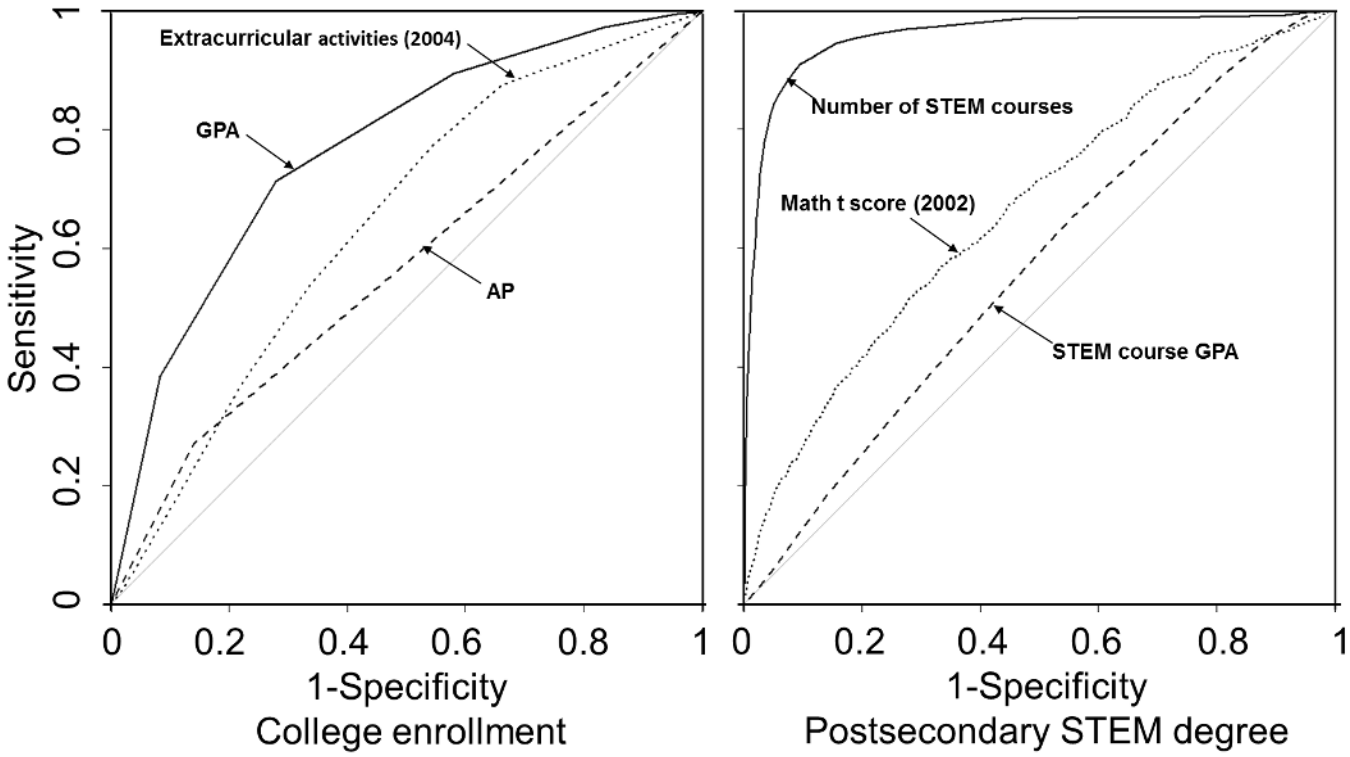
In comparison to the dichotomous predictors in Figure 2A, in Figure 2 Panel B continuous variables are used to predict dropout with ELS:2002 data. Although predictors of "suspension" and "absent" are continuous variables, each with five and four categories respectively, their ROC curves look like dichotomous variables as students in the dataset are suspended or absent on average about once (see Appendix A). As in Figure 2A, the AUC of each curve is provided below the plot of the ROC curves in Figure 2B. The most accurate predictor of dropout in Figure 2B is the Mathematics standardized assessment score with an AUC of 0.722. Table 1 provides the pair-wise AUC comparisons, showing that each predictor is significantly different from each of the others in Figure 2B.

College Enrollment, Postsecondary STEM Degree, and Hard STEM/Soft STEM Occupations

Next, drawing on the prior research noted in the literature review and methods, we provide a series of example applications to provide a range of examples of using ROC AUC across K-12, post-secondary, and occupational prediction outcomes. Our goal is to provide a series of illustrative examples to provide a ready means for K-12 and post-secondary EWI/EWS researchers and practitioners to find accessible examples of the method that may relate to their practice. Figure 3 illustrates ROC curves predicting college enrollment (Figure 3 Panel A) and postsecondary STEM degree attainment (Figure 3 Panel B). Figure 3A compares the accuracy for predicting college enrollment with three variables including overall high school Grade Point Average (GPA), number of extracurricular activities, and if a student ever was enrolled in an Advanced Placement (AP) course. As demonstrated in Figure 3A, overall high school GPA is the most accurate predictor when comparing these three early warning indicators with an AUC of 0.767, which as shown in Table 2 is significantly different than the other two predictors ($p<0.001$).

In Figure 3B we plot the ROC AUC accuracy for predictors of if a student graduates with a postsecondary STEM degree using three continuous variables, number of STEM courses, math t score (2002), and STEM course Grade Point Average (GPA). As demonstrated in Figure 3B, number college STEM courses is the most accurate predictor when comparing these three early

(A) (B)



Predictor:	AUC
GPA:	0.767
Extracurricular. activities (2004):	0.642
AP: 5	0.565

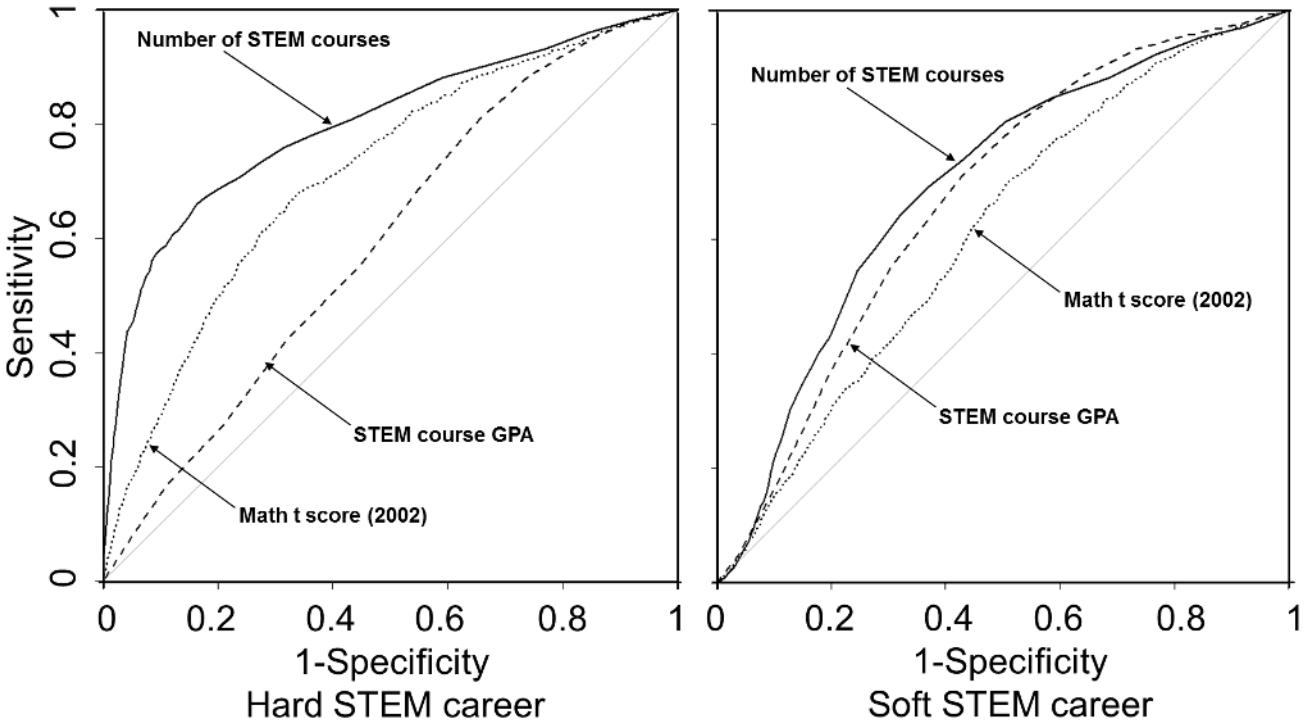
Predictor:	AUC
Number of STEM Courses:	0.957
Math t score (2002):	0.663
STEM course GPA:	0.566

Figure 3: ROC Curves for Predicting College Enrollment and Postsecondary STEM Degree

Table 2: Significance of AUC Difference for College Enrollment Predictors and STEM Career Predictors

Predictor 1	Predictor 2	*Z*	*p*-value
College Enrollment Predictors			
GPA	AP	32.679	<0.001
GPA	Extracurricular activities (2004)	17.070	<0.001
AP	Extracurricular activities (2004)	-10.939	<0.001
Postsecondary STEM Degree Predictors			
Number of STEM Courses	STEM course GPA	39.649	<0.001
Number of STEM Courses	Math t score (2002)	31.625	<0.001
STEM course GPA	Math t score (2002)	-8.712	<0.001

(A) (B)



Predictor:	AUC
Number of STEM Courses:	0.799
Math t score (2002):	0.712
STEM course GPA:	0.593

Predictor:	AUC
Number of STEM Courses:	0.693
STEM course GPA:	0.673
Math t score (2002):	0.611

Figure 4: ROC Curves for Predicting Soft STEM Career and Hard STEM Career

Table 3: Significance of AUC Difference for Soft and Hard STEM Career Predictors

Predictor 1	Predictor 2	*Z*	*p*-value
Soft STEM Career Predictors			
Number of STEM Courses	STEM course GPA	15.055	<0.001
Number of STEM Courses	Math t score (2002)	8.943	<0.001
STEM course GPA	Math t score (2002)	-7.401	<0.001
Hard STEM Career Predictors			
Number of STEM Courses	STEM course GPA	0.846	0.397
Number of STEM Courses	Math t score (2002)	10.523	<0.001
STEM course GPA	Math t score (2002)	11.111	<0.001

warning indicators with an AUC of 0.957, which as shown in Table 2 is significantly different than the other two predictors (p<0.001). Additionally, the R package OptimalCutpoints (López-Ratón et al., 2014) allows us to calculate the optimal cutoff point along the ROC curve of college STEM course number, at 23 (at the apex of the curve), although some students took up to 56 STEM courses in college. This shows that the optimal number of courses in predicting a postsecondary STEM degree is 23 STEM courses. Also, this result shows that in comparison to STEM course grades, the number of STEM courses is more accurate in predicting overall STEM degree completion.

And finally, in Figure 4 we plot the ROC curves predicting hard and soft STEM career occupation at age 26. Figure 4A compares the accuracy for predictors of hard STEM careers of three continuous variables including the Number of STEM Courses, math t score (2002), and college STEM course Grade Point Average (GPA). As demonstrated in Figure 4A, the Number of STEM Courses is the most accurate predictor when comparing these three early warning indicators with an AUC of 0.799, which as shown in Table 3 is significantly different than that of math t score (2002) (p<0.001). In addition, the AUC of STEM course GPA (0.593) is significantly different than that of math t score (2002) (p<0.001).

In Figure 4B we plot the accuracy of predictors of soft STEM careers with three continuous variables, college STEM course number, math t score (2002), and college STEM course Grade Point Average (GPA). In contrast to hard STEM, we show that there is no statistical difference in the accuracy of using either of the predictors of Number of STEM Courses or STEM course GPA in predicting if a student worked in a soft STEM occupation at age 26 (Z=0.846, p=0.397). However, these two predictors are significantly different from the mathematics t-score, each with higher AUC.

DISCUSSION:

The purpose of this study is to provide a means for researchers, policymakers, and practitioners to evaluate and compare the accuracy of educationally relevant predictors of educational outcomes to help inform the EWI/EWS literature through the application of signal detection theory. We demonstrate the applicability of the technique using open source code in R with public data from ELS:2002. Our intention is to help build capacity and inform future EWI/EWS research and applications as having a means to know and correctly compare the accuracy of early warning flags, predictors, and indicators is critical as researchers and educators rely evermore on early warning systems to help identify which students may need additional supports and resources to help them persist and succeed in school. When early warning systems rely on "at risk" indicators that have low accuracy, students are misidentified, resources are expended inefficiently, and many students who could use the resources are missed, at times as many as 40-50% of the students at risk (Bowers et al., 2013). Additionally, it is difficult to improve the indicators and the research for work in the EWI/EWS domain without a rigorous means to assess and compare the accuracy of the predictors, as how are we to know if the predictors and indicators are getting any better without strong accuracy metrics. Our findings in this study show that the application of signal detection theory through ROC AUC and open source R code works well to help identify which indicators are most accurate for an outcome. Our recommendation is that any research or tool that uses a predictor or indicator to identify students at risk provide the ROC plot and AUC data with the significance tests to demonstrate that the indicator under discussion is more accurate than the standard practice as well as similar variables. A strong recent example in this domain is Sullivan, Marr, & Hu (2017), in which the authors used ROC AUC as the criterion to compare logistic regression models and decision tree models in terms of predicting standardized test performance in the Michigan Educational Assessment Program, determining that decision tree models were more accurate with an AUC of 0.92.

Our findings of the accuracy of predictors for high school dropout provide a context to understand results from previous studies. First, our study concurs with Bowers, Sprott, and Taff (2013) as both demonstrate ROC as an accurate diagnostic measure for predictors of dropout. However, whereas Bowers, Sprott, and Taff (2013) use predictor point-estimate positions on the x and y coordinates to compare predictor accuracy, we use AUC to compare predictors, allowing for easy, vivid comparisons. Second, our study provides a new perspective to evaluate predictors of previous studies (e.g., Balfanz et al., 2007). For instance, we show that "absent" has relatively low sensitivity, "any one flag" has low specificity, and "all four flags" and "any three flags" have both low sensitivity and low specificity. The finding that "any one flag" has the highest accuracy of the dichotomous predictors replicates the previous research (Balfanz et al., 2007; Bowers et al., 2013) and demonstrates the utility of an ensemble approach to constructing accurate predictors of dropout using Boolean operators such as "or" versus "and".

Third, our findings indicate that some continuous predictors considered as an accurate means to predict dropout, such as absence and suspension (Adelman, 2006; Allensworth, Nagaoka, & Johnson, 2018; Balfanz et al., 2007; Bowers, & Sprott, 2012a, 2012b; Kemple et al., 2013; Quadri & Kalyankar, 2010; Soland, 2017), actually have low accuracy in comparison to the other predictors. For these continuous predictors of high school dropout, it is interesting that grade 10 mathematics and reading standardized test scores were the two most accurate predictors, with mathematics having the highest accuracy. Across the research on dropout flags (Bowers et al., 2013), standardized test scores have not received a large amount of attention (Bowers, 2010; Kemple et al., 2013; Soland, 2013; Soland, 2017), as previously authors have focused on a wide constellation of dropout predictors, having previously lacked a useful means to compare the accuracy of predictors (see literature review). However, here, the NCES ELS:2002 standardized mathematics and reading scores are test scores equated to the national NAEP (National Assessment of Educational Progress) standardized tests (Ingles et al., 2014).

NAEP, known colloquially as "the nation's report card" is well-known to be a difficult assessment in reading and mathematics, and that state standardized test scores have a poor history of alignment with NAEP (Bandeira de Mello, Bohrnstedt, Blankenship, & Sherman, 2015). Thus, the finding that the ELS:2002 grade 10 mathematics and reading test scores perform well in comparison to the other continuous dropout flags tested here does not necessarily mean that most or any other state standardized test score will perform in the same way. We encourage future research in this area to further include standardized subject assessments in dropout flag and predictor research.

Our findings of the accuracy of college enrollment predictors in terms of ROC AUC provide a means to identify differences between the accuracy of these predictors. To date, as noted in the literature review above, a central problem in this research domain is that researchers tend to depend on regression outcomes and the statistical significance of individual coefficients, rather than on magnitude of effect, to make claims that a variable "significantly predicts" an outcome. Indeed, using this style of a regression coefficient framework, researchers are hampered in their ability to assess the differences in predictor accuracy for variables such as GPA, AP, and extracurricular activities, as all three predictors are significant in predicting if high school students will enter college. Nevertheless, from the perspective of signal detection theory and sensitivity and specificity, we find that these three predictors do have significant differences in their accuracy in predicting college enrollment. For instance, high school GPA was significantly more accurate in predicting college enrollment than students having taken Advanced Placement courses (AP), with extracurricular activities lying between these two in accuracy. The point that AP courses are the lowest of these three makes sense, as given that AP is billed as preparation for college, taking AP courses may not particularly differentiate between students enrolling in post-secondary institutions (Ackerman, Kanfer, & Calderwood, 2013; Sublett & Gottfried, 2017). That high school GPA performs well in the accuracy of predicting college enrollment corresponds with the research on grades and GPA (Bowers, 2009; Bowers, 2011; Bowers, 2019; Brookhart et al., 2016) which indicates that grades represent a valid assessment of engaged participation in the education system, and thus present a strong signaling effect for advancing through the system or not (Pattison, Grodsky, & Muller, 2013).

Likewise, our findings of using ROC AUC to examine the accuracy of predictors for postsecondary STEM degrees, hard STEM careers, and soft STEM careers allow for comparing predictors from the perspective of sensitivity and specificity. First, we identify Number of STEM Courses as a highly accurate predictor for postsecondary STEM degrees, with an AUC above 0.9. While this result makes sense, as the total number of STEM courses is strongly related to the requirements to graduate with a STEM degree in college, the point we show here is that it is the total number of STEM courses that is highly accurate, in comparison to STEM GPA, and that the optimal number of STEM courses students need to graduate from a postsecondary institution with a STEM degree is 23, aligning with previous research (Chen, 2013). Additionally, this finding aligns well with the previous research in postsecondary STEM degree attainment that has highlighted the number of STEM courses as a central component of successful STEM degree attainment (Crisp, Nora, & Taggart, 2009; Engberg & Wolniak, 2013).

Second, for predicting a hard STEM career or a soft STEM career at age 26, we demonstrate interesting differences in the accuracy of the predictors, finding that the total number of courses is the most accurate predictor for hard STEM, but that there is no statistical difference in the accuracy of the number of STEM courses versus STEM GPA for predicting soft STEM careers. Additionally, while grade 10 mathematics standardized test score sits between number of STEM courses and STEM course GPA for predicting hard STEM careers, mathematics standardized test score has the lowest accuracy of these three predictors for a soft STEM career. This work extends the previous research on STEM college experiences that lead to careers that has shown that number of STEM courses, STEM GPA, and mathematics scores are all related to STEM career outcomes (Chen, 2013; Sithole et al., 2017). Additionally, it is interesting to note the STEM course GPA AUC between hard STEM and soft STEM, as studies have shown that for STEM overall in college, students receive lower GPAs overall than non-STEM students (Chen, 2013; Sithole et al., 2017). Overall, our results may suggest that while number of STEM courses in college is an important variable in predicting STEM career outcomes, there may be significant differences for STEM course GPA and mathematics scores between the hard STEM and soft STEM disciplines. Thus, overall, while we provide these findings as illustrative examples for the application of ROC AUC to EWI/EWS, the reasons for the differences in the accuracy of the predictors is of further interest, but outside the scope of the present study. We encourage future research in these areas.

Limitations:

While we believe our results are rigorous and applicable, this study is limited in three main ways. First, we note that the *auc* function of pROC gives the AUC of a predictor, but sometimes this AUC is slightly different from the one from the function *roc.test* where the AUC of this predictor is being compared with the AUC of another predictor. The deviation may be as large as 0.05, unsubstantial in practice, and we have not found researchers to have mentioned this point elsewhere. This is most likely due to Delong et al.'s trapezoid rule dividing the AUC differently when calculating the AUC of a single predictor from when calculating the AUCs of two correlated predictors. Second, since pROC does listwise deletion on both the predictor and the outcome variable, the sample sizes of the ELS:2002 dataset for the outcomes listed in Appendix A were shrunk for calculations of the AUC, making the results of the present study generalizable to only the data included in the analyses. In the future, we encourage researchers to delve more deeply into the missing values issue, such as multiple

imputations, before using pROC to calculate AUC so that the results will be more generalizable. As we were attempting to provide an example in practice with data similar to what a school district may have, this type of missing data imputation is outside the scope of the present study. Third, the generalizability of the present student is limited as we refrained from using the probabilistic weights of the ELS:2002 dataset in either calculating AUC's or comparing the AUC's of different predictors. Thus, the current study should be considered as a descriptive rather than an inferential analysis. Again, we encourage future researchers to take our study as a starting point for ROC AUC analysis and work to incorporate normalized weights on data similar to ELS:2002 data to compute AUC.

Practical Implications and Conclusion:

Our findings have three main practical implications. First, with the need to report accuracy of outcomes, education researchers can calculate AUC with R packages such as pROC. Practitioners and analysts can build models on the data in institutions and build ROC plots and calculate AUC for each predictor, with the R code provided here for reference (see Online Appendix S1 http://doi.org/10.7916/D8K94RDD). Second, ROC AUC allows policymakers to check and confirm the accuracy of predictors currently in use to make decisions on which at-risk predictors are the most accurate to predict important educational outcomes, and which predictors have low accuracy. Third, ROC AUC helps policymakers to identify accurate EWIs for the outcomes that are most important to their communities such that policies may cover the largest possible proportion of students at risk through using the most accurate predictors and maximizing the "hits" while minimizing the "false-alarms". For education administrators and policymakers, this type of accuracy information is crucial in their work to provide the limited resources of the education system to help support specific student needs.

Overall, we believe that the present research contributes to promoting AUC as an accurate measure to evaluate predictors of Early Warning Systems (EWS) and Early Warning Indictors (EWI), having shown the advantages of AUC in comparing the accuracy of predictors of different education outcomes. The findings of this study suggest that in the future, education researchers should 1) calculate the AUC of each predictor to give a baseline indication of the predictor's accuracy; and 2) carry out pairwise significance tests on the AUCs of different predictors to show if the predictors are significantly different in terms of accuracy.

Suggested Citation:

Bowers, A.J., Zhou, X. (2019) Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. *Journal of Education for Students Placed At Risk*, 24(1) 20-46. https://doi.org/10.1080/10824669.2018.1523734

REFERENCES:

Adelman, C. (2006). The toolbox revisited: Paths to degree completion from high school through college. Washington, DC: US Department of Education. Retrieved from: http://files.eric.ed.gov/fulltext/ED490195.pdf

Agasisti, T., Bowers, A.J. (2017). Data Analytics and Decision-Making in Education: Towards the Educational Data Scientist as a Key Actor in Schools and Higher Education Institutions. In Johnes, G., Johnes, J., Agasisti, T., López-Torres, L. (Eds.) *Handbook of Contemporary Education Economics* (184-210). Cheltenham, UK: Edward Elgar Publishing. https://doi.org/10.7916/D8PR95T2

Allensworth, E. M. (2006). From High School to the Future: A First Look at Chicago Public School Graduates' College Enrollment, College Preparation, and Graduation from Four-Year Colleges. *Consortium on Chicago School Research*. http://files.eric.ed.gov/fulltext/ED499368.pdf

Allensworth, E. M. (2013). The use of ninth-grade early warning indicators to improve Chicago schools. *Journal of Education for Students Placed at Risk (JESPAR)*, *18*(1), 68-83. DOI: 10.1080/10824669.2013.745181

Allensworth, E. M., & Easton, J. Q. (2007). What Matters for Staying On-Track and Graduating in Chicago Public Highs Schools: A Close Look at Course Grades, Failures, and Attendance in the Freshman Year. Research Report. *Consortium on Chicago School Research*. http://files.eric.ed.gov/fulltext/ED498350.pdf

Allensworth, E. M., Gwynne, J. A., Moore, P., & de la Torre, M. (2014). Looking Forward to High School and College: Middle Grade Indicators of Readiness in Chicago Public Schools. *University of Chicago Consortium on Chicago School Research*. http://files.eric.ed.gov/fulltext/ED553149.pdf

Allensworth, E., & Luppescu, S. (2018). Why do students get good grades, or bad ones? The influence of the teacher, class, school, and student Retrieved from Chicago: IL: https://consortium.uchicago.edu/sites/default/files/publications/Why%20Do%20Students%20Get-Apr2018-Consortium.pdf

Allensworth, E.M., Nagaoka, J., & Johnson, D.W. (2018). *High school graduation and college readiness indicator systems: What we know, what we need to know*. Chicago, IL: University of Chicago Consortium on School Research.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387-415. DOI: 10.1016/0022-2496(75)90001-2

Baker, R. S. (2013). Learning, Schooling, and Data Analytics. In M. Murphy, S. Redding, & J. Twyman (Eds.), Handbook on innovations in learning Philadelphia, PA: Center on Innovations in Learning, Temple University; Charlotte, NC: Information Age Publishing. Retrieved from http://www.centeril.org/

Baker, R. S. (2014). Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent systems, 29*(3), 78-82. DOI: 10.1109/MIS.2014.42

Baker, R.S. (2015). *Big Data and Education*. 2nd Edition. New York, NY: Teachers College, Columbia University.

Balfanz, R., & Boccanfuso, C. (2007). Falling off the path to graduation: Middle grade indicators in [an unidentified northeastern city]. Baltimore, MD: Center for Social Organization of Schools.

Balfanz, R., Herzog, L., & Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, *42*(4), 223-235. http://www.tandfonline.com/doi/pdf/10.1080/00461520701621079?needAccess=true. DOI: 10.1080/00461520701621079

Bandeira de Mello, V., Bohrnstedt, G., Blankenship, C., & Sherman, D. (2015). Mapping State Proficiency Standards Onto NAEP Scales: Results From the 2013 NAEP Reading and Mathematics Assessments (NCES 2015-046) (NCES 2015-046). Retrieved from Washington, DC: https://files.eric.ed.gov/fulltext/ED557749.pdf

Becker, J., Schools, P. P., Levinger, B., Schools, P. G. S. C. P., Sims, A., & Whittington, A. (2014). Student Success and College Readiness: Translating Predictive Analytics Into Action. http://sdp.cepr.harvard.edu/files/cepr-sdp/files/sdp-fellowship-capstone-student-success-college-readiness.pdf

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2009). *Crossing the finish line: Completing college at America's public universities*. Princeton University Press.

Bowers, A.J. (2019) Report Card Grades and Educational Outcomes. In Guskey, T., Brookhart, S. (Eds.) What We Know About Grading: What Works, What Doesn't, and What's Next, (p.32-56). Alexandria, VA: Association for Supervision and Curriculum Development.

Bowers, A.J. (2017) Quantitative Research Methods Training in Education Leadership and Administration Preparation Programs as Disciplined Inquiry for Building School Improvement Capacity. *Journal of Research on Leadership Education*, 12(1), p. 72-96. DOI: 10.1177/1942775116659462

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation*, 17(3), 141-159. doi:10.1080/13803611.2011.597112

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades,data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment Research and Evaluation, 15*(7), 1–18. http://www.pareonline.net/pdf/v15n7.pdf

Bowers, A. J. (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration*, 47(5), 609-629. doi:10.1108/09578230910981080

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, *96*(2), 77-100. https://muse.jhu.edu/article/496998/pdf

Bowers, A.J., Sprott, R. (2012a) Why Tenth Graders Fail to Finish High School: A Dropout Typology Latent Class Analysis. *The Journal of Education for Students Placed at Risk* (JESPAR), 17(3), 129-148. DOI: 10.1080/10824669.2012.692071

Bowers, A.J., Sprott, R. (2012b) Examining the Multiple Trajectories Associated with Dropping Out of High School: A Growth Mixture Model Analysis. The Journal of Educational Research, 105(3), 176-195. DOI: 10.1080/00220671.2011.552075

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., . . . Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of Educational Research*, 86(4), 803-848. doi:10.3102/0034654316672069

Carlson, S. E. (2018). Identifying Students at Risk of Dropping out: Indicators and Thresholds Using ROC Analysis. http://digitalcommons.georgefox.edu/edd/114

Chen, X. (2013). STEM Attrition: College Students' Paths into and out of STEM Fields. Statistical Analysis Report. NCES 2014-001. *National Center for Education Statistics*.

Clark Blickenstaff, (2005). Women and science careers: leaky pipeline or gender filter?. *Gender and Education*, *17*(4), 369-386. DOI: 10.1080/09540250500145072

Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution. *American Educational Research Journal, 46*(4), 924-942. DOI: 10.3102/0002831209349460

Cummings, K. D., & Smolkowski, K. (2015). Selecting students at risk of academic difficulties. *Assessment for Effective Intervention, 41*(1), 55-61. DOI: 10.1177/1534508415590396

D'Agostino, J. V., Rodgers, E., & Mauck, S. (2018). Addressing Inadequacies of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly*, 53(1), 51-69. doi:10.1002/rrq.181

Davis, M., Herzog, L., & Legters, N. (2013). Organizing Schools to Address Early Warning Indicators (EWIs): Common Practices and Challenges. Journal of Education for Students Placed at Risk (JESPAR), 18(1), 84-100. DOI: 10.1080/10824669.2013.745210

DeLong, E., DeLong, D., & Clarke-Pearson, D. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics, 44*(3), 837-845. DOI: 10.2307/2531595. http://www.jstor.org/stable/2531595

Dutta-Moscato, J., Gopalakrishnan, V., Lotze, M. T., & Becich, M. J. (2014). Creating a pipeline of talent for informatics: STEM initiative for high school students in computer science, biology, and biomedical informatics. *Journal of Pathology Informatics*, *5*. DOI: 10.4103/2153-3539.129448

Engberg, M. E., & Wolniak, G. C. (2013). College student pathways to the STEM disciplines. *Teachers College Record*, 115(1), 1-27.

Fanjul-Hevia, A., & González-Manteiga, W. (2018). A comparative study of methods for testing the equality of two or more ROC curves. *Computational Statistics*, 1-21. DOI: 10.1007/s00180-017-0783-6

Faria, A. M., Sorensen, N., Heppen, J., Bowdon, J., Taylor, S., Eisner, R., & Foster, S. (2017). Getting Students on Track for Graduation: Impacts of the Early Warning Intervention and Monitoring System after One Year. REL 2017-272. *Regional Educational Laboratory Midwest*. https://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=388

Fogarty, J., Baker, R. S., & Hudson, S. E. (2005, May). Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005* (pp. 129-136). Canadian Human-Computer Communications Society.

Frazelle, S., & Nagel, A. (2015). A practitioner's guide to implementing early warning systems. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northwest.

Geiser, S., & Santelices, M. V. (2007). Validity of High-School Grades in Predicting Student Success beyond the Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes. Research & Occasional Paper Series: CSHE. 6.07. *Center for studies in higher education*. http://files.eric.ed.gov/fulltext/ED502858.pdf

Gleason, P., & Dynarski, M. (2002). Do we know whom to serve? Issues in using risk factors to identify dropouts. *Journal of Education for Students Placed At Risk*, *7*(1), 25-41. DOI: 10.1207/S15327671ESPR0701_3

Gönen, M. (2007). Analyzing Receiver Operating Characteristic Curves With SAS. Cary, NC: SAS Publishing.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36. DOI: 10.1148/radiology.143.1.7063747

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*(3), 839-843. DOI: 10.1148/radiology.148.3.6878708.

Horn, A. S., & Lee, G. (2017). Evaluating the Accuracy of Productivity Indicators in Performance Funding Models. *Educational Policy*. DOI: 10.1177/0895904817719521

Hughes, J., & Petscher, Y. (2016). A guide to developing and evaluating a college readiness screener. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Ikbal, S., Tamhane, A., Sengupta, B., Chetlur, M., Ghosh, S., & Appleton, J. (2015). On early prediction of risks in academic performance for students. IBM Journal of Research and Development, 59(6), 5:1-5:14. DOI: 10.1147/JRD.2015.2458631

Ingels, S. J., Pratt, D. J., Alexander, C. P., Jewell, D. M., Lauff, E., Mattox, T. L., . . . Christopher, E. (2014). Education Longitudinal Study of 2002 (ELS:2002) Third Follow-Up Data File Documentation (NCES 2014-364). Retrieved from Washington, DC: http://nces.ed.gov/pubs2014/2014364.pdf

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013, September). Facing imbalanced data--recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (pp. 245-251). IEEE. DOI: 10.1109/ACII.2013.47

Johnson, E., & Semmelroth, C. (2010). The Predictive Validity of the Early Warning System Tool. *NASSP Bulletin, 9*4(2), 120-134. DOI: 10.1177/0192636510380789

Jordan, N., Glutting, J., Ramineni, C., & Watkins, M. (2010). Validating a Number Sense Screening Tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review, 39*(2), 181-195. Retrieved from http://ezproxy.cul.columbia.edu/login?url=https://search.proquest.com/docview/608709943?accountid=10226

Kemple, J. J., Segeritz, M. D., & Stephenson, N. (2013). Building On-Track Indicators for High School Graduation and College Readiness: Evidence from New York City. *Journal of Education for Students Placed at Risk (JESPAR), 18*(1), 7-28. DOI: 10.1080/10824669.2013.747945

Knight, S., & Shum, S. B. (2018). Theory and Learning Analytics. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), Handbook of Learning Analytics (pp. 17-22): Society for Learning Analytics Research.

Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *JEDM-Journal of Educational Data Mining*, *7*(3), 18-67. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/JEDM082

Koon, S., & Petscher, Y. (2016). Can scores on an interim high school reading assessment accurately predict low performance on college readiness exams? . Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Krumm, A. E., Means, B., & Bienkowski, M. (2018). *Learning Analytics Goes to School: A Collaborative Approach to Improving Education*. New York: Routledge.

Lacefield, W. E., & Applegate, E. B. (2018). Data Visualization in Public Education: Longitudinal Student-, Intervention-, School-, and District-Level Performance Modeling. Paper presented at the Annual meeting of the American Educational Research Association, New York, NY.

Lacefield, W., Applegate, B., Zeller, P. J., & Carpenter, S. (2012). Tracking students' academic progress in data rich

but analytically poor environments. Paper presented at the Annual meeting of the American Educational Research Association, Vancouver, BC.

Langenbucher, J., Labouvie, E., & Morgenstern, J. (1996). Measuring diagnostic agreement. *Journal of Consulting and Clinical Psychology*, 64(6), 1285. DOI: 10.1037/0022-006X.64.6.1285

Laracy, S. D., Hojnoski, R. L., & Dever, B. V. (2016). Assessing the classification accuracy of early numeracy curriculum-based measures using receiver operating characteristic curve analysis. *Assessment for Effective Intervention, 41*(3), 172-183. DOI: 10.1177/1534508415621542

Liao, H. F., Yao, G., Chien, C. C., Cheng, L. Y., & Hsieh, W. S. (2014). Likelihood ratios of multiple cutoff points of the Taipei City Developmental Checklist for Preschoolers, 2nd version. *Journal of the Formosan Medical Association, 113*(3), 179-186. DOI: 10.1016/j.jfma.2011.10.55

López-Ratón, M., Rodríguez-Álvarez, M. X., Suarez, C. C., & Sampedro, F. G. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software, 61*(8), 1-36.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers & Education, 54(2), 588-599. DOI: 10.1016/j.compedu.2009.09.008

Mahoney, J. L., & Cairns, R. B. (1997). Do extracurricular activities protect against early school dropout?. *Developmental psychology*, *33*(2), 241. DOI: 10.1037/0012-1649.33.2.241

National Center on Response to Intervention. (2010). Users guide to universal screening tools chart. Washington, DC: National Center on Response to Intervention, Office of Special Education Programs, U.S. Department of Education. Retrieved from http://www.rti4success.org/sites/default/files/UniversalScreeningUsersGuide.pdf

Neild, R. C., Balfanz, R., & Herzog, L. (2007). An early warning system. Educational Leadership, 65(2), 28–33. Retrieved from http://new.every1graduates.org/wp-content/uploads/2012/03/Early_Warning_System_Neild_Balfanz_Herzog.pdf

Nicholls, G., Wolfe, H., Besterfield-Sacre, M., & Shuman, L. (2010). Predicting STEM degree outcomes based on eighth grade data and standard test scores. *Journal of Engineering Education, 99*(3), 209-223. DOI: 10.1002/j.2168-9830.2010.tb01057.x

Noble, J. P., & Sawyer, R. L. (2004). Is high school GPA better than admission test scores for predicting academic success in college? *College and University, 79*(4), 17–22.

Ocumpaugh, J., Baker, R. S., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology, 45*(3), 487-501. DOI: 10.1111/bjet.12156

Ocumpaugh, J., Baker, R. S., San Pedro, M. O., Hawn, M. A., Heffernan, C., Heffernan, N., & Slater, S. A. (2017, March). Guidance counselor reports of the ASSISTments college prediction model (ACPM). In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 479-488). ACM. DOI: 10.1145/3027385.3027435

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk assessment with young offenders: A meta-analysis of three assessment measures. Criminal Justice and Behavior, 36(4), 329-353. DOI: 10.1177/009385480933145

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the Sky Falling? Grade Inflation and the Signaling Power of Grades. *Educational Researcher*, 42(5), 259-265. doi:10.3102/0013189x13481382

Pelánek, R. (2014). A brief overview of metrics for evaluation of student models. In *Workshop Approaching Twenty Years of Knowledge Tracing (BKT20y)* (p. 6).

Pelánek, R. (2015). Metrics for evaluation of student models. *JEDM-Journal of Educational Data Mining*, *7*(2), 1-19.

Perna, L., Lundy-Wagner, V., Drezner, N. D., Gasman, M., Yoon, S., Bose, E., & Gary, S. (2009). The contribution of HBCUs to the preparation of African American women for STEM careers: A case study. *Research in Higher Education*, *50*(1), 1-23. DOI: 10.1007/s11162-008-9110-y

Phillips, M., Yamashiro, K., Farrukh, A., Lim, C., Hayes, K., Wagner, N., . . . Chen, H. (2015). Using Research to Improve College Readiness: A Research Partnership Between the Los Angeles Unified School District and the Los Angeles Education Research Institute. *Journal of Education for Students Placed at Risk, 20*(1), 141-168. DOI: 10.1080/10824669.2014.990562

Piety, P. J., Hickey, D. T., & Bishop, M. (2014). Educational data sciences: Framing emergent practices for analytics of learning, organizations, and systems. Paper presented at the Proceedings of the Fourth International Conference on Learning Analytics and Knowledge. DOI: 10.1145/2567574.2567582

Quadri, M. M., & Kalyankar, N. V. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology, 10*(2), 2-5. Retrieved from https://computerresearch.org/index.php/computer/article/view/891

R Development Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Renzulli, J. S., & Park, S. (2000). Gifted dropouts: The who and the why.*Gifted Child Quarterly*, *44*(4), 261-271. DOI: 10.1177/001698620004400407

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. Law and Human Behavior, 29(5), 615-620. DOI: 10.1007/s10979-005-6832-

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 77. DOI: 10.1186/1471-2105-12-77

Robnett, R. D., & Leaper, C. (2013). Friendship groups, personal motivation, and gender in relation to high school students' STEM career interest. *Journal of Research on Adolescence*, *23*(4), 652-664. DOI: 10.1111/jora.12013

Rumberger, R. W., Addis, H., Allensworth, E., Balfanz, R., Duardo, D., Dynarksi, M., . . . Tuttle, C. (2017). Preventing dropout in secondary schools (NCEE 2017-4028). Retrieved from Washington, DC: https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/wwc_dropout_092617.pdf

Siemens, G. (2013). Learning Analytics: The Emergence of a Discipline. *American Behavioral Scientist,* 57(10), 1380-1400. DOI: 10.1177/0002764213498851

Sithole, A., Chiyaka, E. T., McCarthy, P., Mupinga, D. M., Bucklein, B. K., & Kibirige, J. (2017). Student Attraction, Persistence and Retention in STEM Programs: Successes and Continuing Challenges. *Higher Education Studies*, 7(1), 46-59.

Soland, J. (2013). Predicting high school graduation and college enrollment: Comparing early warning indicator data and teacher intuition. *Journal of Education for Students Placed at Risk (JESPAR)*, *18*(3-4), 233-262. DOI: 10.1080/10824669.2013.833047

Soland, J. (2017). Combining Academic, Noncognitive, and College Knowledge Measures to Identify Students Not on Track For College: A Data-Driven Approach. *Research & Practice in Assessment*, *12*.

Stuit, D., O'Cummings, M., Norbury, H., Heppen, J., Dhillon, S., Lindsay, J., & Zhu, B. (2016). *Identifying early warning indicators in three Ohio school Districts* (REL 2016- 118). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved from http://ies.ed.gov/ncee/edlabs

Sullivan, W., Marr, J., & Hu, G. (2017). A Predictive Model for Standardized Test Performance in Michigan Schools. In *Applied Computing and Information Technology* (pp. 31-46). Springer, Cham.

Supovitz, J. A., Foley, E., & Mishook, J. (2012). In search of leading indicators in education. *Education Policy Analysis Archives, 20*(19), 1-23.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems.*Science*, *240*(4857), 1285-1293. DOI: 10.1126/science.3287615

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest*, 1(1), 1-26. doi:10.1111/1529-1006.001

Swets, J. A., & Pickett, R. M. (1982). Evaluation of diagnostic systems: methods from signal detection theory. *Cognition*.

Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M., & Appleton, J. (2014). Predicting student risks through longitudinal analysis. Paper presented at the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, New York, USA. http://dl.acm.org/citation.cfm?id=2623355. DOI: 10.1145/2623330.2623355

Torres, D. D., Bancroft, A., & Stroub, K. (2015). Evaluating High School Dropout Indicators and Assessing Their Strength.

U.S. Department of Education. (2014, March 26). *Academic Competitiveness Grants and National Science and Mathematics Access to Retain Talent (SMART) Grants*. Retrieved from https://www2.ed.gov/programs/smart/index.html

Van Inwegen, E. G., Wang, Y., Adjei, S., & Heffernan, N. (n.d.). An Examination of Metrics that Describe User Models.

Vivo, J. M., & Franco, M. (2008). How does one assess the accuracy of academic success predictors? ROC analysis applied to university entrance factors. *International Journal of Mathematical Education in Science and Technology, 39*(3), 325-340. DOI: 10.1080/00207390701691566

Whittaker, R. C. (2014). Teaching in context using a mobile phone scenario. Master thesis, University of Salford.

Willis. (2013). Higher education in science, technology, engineering and mathematics: Science and Technology Committee Report — Motion to Take Note— in the House of Lords at 5:21 pm on 21st March 2013.

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing, 27*, 11-23. DOI: 10.1016/j.asw.2015.06.003

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

About the Authors:

Alex J. Bowers is an associate professor of Education Leadership at Teachers College, Columbia University. His research interests include organizational level data analytics, organizational behavior, school and district leadership, data-driven decision making, high school dropouts and completion, educational assessment and accountability, education technology and school facilities financing.
ORCID: 0000-0002-5140-6428

Xiaoliang Zhou is a PhD candidate of Measurement, Evaluation, and Statistics at Teachers College, Columbia University. His research interests include cognitive diagnostic modeling, item response theory, hierarchical modeling, rater effect models, structural equation modeling, data mining, machine learning, data visualization, natural language processing, and deep learning. ORCID: 0000-0003-0467-6123

APPENDIX:

Appendix A: Variable Descriptive Statistics and Labels

Variable Name	N	M	SD	Min.	Max.	ELS:2002 Variable
Dropout	16,197	0.113	0.317	0	1	F2EVERDO; 1=Evidence of a dropout episode
Absent	12,286	0.130	0.336	0	1	BYP52E; 1=School contacted parent about poor attendance 1- 4 times
Suspension	14,476	0.081	0.272	0	1	BYS24F; 1=Suspended/put on probation 1- 5 times
Misbehavior	12,457	0.072	0.259	0	1	BYP51; 1= Ever had behavior problem at school
1st quantile Math t score (2002)	15,892	0.250	0.433	0	1	BYTXMSTD; 1=Math test standardized score below 1st quantile
One or more flags	16,197	0.371	0.483	0	1	1= One or more flags (absent, lowest quantile reading score, suspension, misbehavior)
Any one flag	13,506	0.445	0.497	0	1	1= Any one flag
Any two flags	12,381	0.124	0.330	0	1	1= Any two flags
Any three flags	13,949	0.025	0.155	0	1	1= Any three flags
All four flags	15,580	0.004	0.059	0	1	1= All four flags
Extracurricular activities (2002)	14,446	4.773	5.700	0	21	BYS42; Hours/week spent on extracurricular activities
Math t score (2002)	15,892	50.710	9.912	19.380	86.680	BYTXMSTD; Math test standardized score
Reading t score (2002)	15,892	50.526	9.885	22.570	78.760	BYTXRSTD; Reading test standardized score
College enrollment	10,511	0.546	0.498	0	1	F2PS0601; 1= Enrolled in a 4-yr institution
GPA	14,796	3.912	1.543	0	6	F1RGPP2; GPA for all courses taken in the 9th - 12th grades
Extracurricular activities (2004)	14,073	3.182	1.898	1	8	F1S27; Hours/week spent on extracurricular activities
AP	14,368	0.182	0.386	0	1	BYS33A; 1= Ever in Advanced Placement program
Postsecondary STEM degree	6,936	0.167	0.373	0	1	F3TZSTEM1CRED; 1= Ever earned a postsecondary credential in a STEM field as of June 2013 (SMART grant definition)
Number of STEM Courses	11,540	9.257	10.396	0	56	F3TZSTEM1TOT; Transcript: Number of known STEM courses taken (using SMART Grant definition of STEM)
STEM course GPA	10,755	2.586	0.938	0	4	F3TZSTEM2GPA; Transcript: GPA for all known STEM courses (using NSF definition of science, engineering, and related fields)
Hard STEM career	12,796	0.063	0.243	0	1	F3STEMOCCCUR; 1= Life and Physical Science, Engineering, Mathematics, and Information Technology Occupations
Soft STEM career	12,796	0.078	0.268	0	1	F3STEMOCCCUR; 1=Social Science Occupations/ Health Occupations

Appendix B: Significance of AUC Difference for Balfanz et al. (2007) Dropout Predictors

	Absent	1st quantile Math t score (2002)	Suspension	Misbehavior	One or more flags	Any one flag	Any two flags	Any three flags
1st quantile Math t score	0.330							
Suspension	<0.001	<0.001						
Misbehavior	<0.001	<0.001	0.719					
One or more flags	<0.001	<0.001	<0.001	<0.001				
Any one flag	<0.001	<0.001	<0.001	<0.001	1.000			
Any two flags	<0.001	0.089	<0.001	<0.001	<0.001	<0.001		
Any three flags	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	
All four flags	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001