Annals of Dyslexia (2022) 72:445–460
https://doi.org/10.1007/s11881-022-00261-5

Screening screeners: calculating classification indices using correlations and cut-points

Ashley A. Edwards[1] · Wilhelmina van Dijk[1] · Christine M. White[1] · Christopher Schatschneider[1]

Received: 19 January 2021 / Accepted: 16 May 2022 / Published online: 10 June 2022

Abstract
Given the recent push for universal screening, it is important to take into account how well a screener identifies children at risk for reading problems as well as how screener and sample information contribute to this classification. Picking the best cut-point for a particular sample and screening goal can be challenging given that test manuals often report classification information for a specific cut-point and sample base rate which may not generalize to other samples. By assuming a bivariate normal distribution, it is possible to calculate all of the classification information for a screener based on the correlation between the screener and outcome, the cut-point on the outcome (i.e., the base rate in the sample), and the cut-point on the screener. We provide an example with empirical data to validate these estimation procedures. This information is the basis for a free online tool that provides classification information for a given correlation between screener and outcome and cut-points on each. Results show that the correlation between screener and outcome needs to be greater than .9 (higher than observed in practice) to obtain good classification. These findings are important for researchers, administrators, and practitioners because current screeners do not meet these requirements. Since a correlation is dependent on the reliability of the measures involved, we need screeners with better reliability and/or multiple measures to increase reliability. Additionally, we demonstrate the impact of base rate on positive predictive power and discuss how gated screening can be useful in samples with low base rates.

Keywords Classification · Identification · Screening

In 2004, the reauthorization of the Individuals with Disabilities Education Act (IDEA) included the provision that eligibility for special education services under the learning disabilities category could be determined using a response-to-intervention (RTI) process. The focus on RTI and the requirement that states actively find children who are suspected to have a disability even if they have not been identified and are progressing through grades

✉ Ashley A. Edwards
aedwards@psy.fsu.edu

[1] Department of Psychology, Florida State University, 1107 West Call Street, Tallahassee, FL 32306, USA

(IDEA, 2012), coupled with the focus on Reading First through the No Child Left Behind Act of 2001 (NCLB, 2008), may have been partly responsible for an increase in universal screening policies in individual states' education laws specifically intended to identify children at risk for dyslexia (Gearin et al., 2018).

The landscape around laws for dyslexia is rapidly changing. Youman and Mather (2018) stated that as of March 2018, 42 US states had laws specifically addressing dyslexia, with a subset of 18 states including provisions on universal screening methods for reading development (Youman & Mather, 2018). Later that same year, Gearin and colleagues (Gearin et al., 2018) reported 20 states with laws for screening. In January 2022, the National Center on Improving Literacy website showed 38 states with screening provisions in their laws (National Center on Improving Literacy, n.d.).

Research on screening

The increase in state legislation on dyslexia, particularly the emphasis on universal screening measures to promote early identification of students at risk for dyslexia, has nourished a renewed impetus for research on screening. Part of this research focuses on actively developing specific screeners for dyslexia (e.g., the Boston Early Literacy Screener [BELS] and the Application for Readiness in Schools and Learning Evaluation [AppRISE]). Other teams of researchers are identifying existing reading skill measures used in schools that have adequate predictive validity for reading outcomes and determining appropriate cutoff scores that will lead to the greatest accuracy in prediction (e.g., Kent et al., 2019; Thomas & January, 2021).

Additionally, since recent work on multiple-deficit models of dyslexia has indicated that screeners for early identification should focus on two or more factors during assessment (Catts & Petscher, 2018), several initiatives have specifically focused on developing risk indicators. One of these initiatives is the Earlier Assessment for Reading Success (EARS), an online calculator of reading comprehension and language difficulty. The calculator uses results from one or more K-3 curriculum-based measures to predict the probability of difficulties (Petscher et al., 2016). An earlier version of a similar calculator (Catts et al., 2001) used results of measures in five potential areas of concern (phonological awareness, rapid naming, letter identification, sentence repetition, and mother's education) to predict risk status. Unfortunately, this calculator did not have the necessary accuracy in predicting risk (Catts & Petscher, 2018).

A different part of screening research focuses on identification of students at risk of not reaching proficiency on state assessments. Thomas and January (2021) examined two alternative tests (the Measure of Academic Progress—a computer adaptive test—and the Strategic Teaching and Evaluation of Progress—a developmental reading assessment) as potential screeners to predict proficiency on state reading assessments in 3rd grade. While the Measure of Academic Progress performed better than the Strategic Teaching and Evaluation of Progress, both measures had insufficient classification accuracy, with many students identified as not at risk who later failed the state assessments. In a study comparing the predictive accuracy of curriculum-based (screening) measures and early school drop-out warning signs (e.g., office discipline referrals, attendance rate, and course failures) to proficiency on a state reading assessment in middle school, Stevenson (2017) found a combination of the previous year's state reading assessment scores and office discipline referrals classified students more accurately than the curriculum-based (screening) measures in both 7th and 8th grades.

Why check for accuracy?

It is important to validate the accuracy of screening measures for two reasons. First, inaccurate measures could lead to over-identification of students possibly at risk for reading difficulties through the inclusion of false positives. This is potentially harmful to school environments because they may allocate additional resources and time to students who in fact do not require reading interventions (Jenkins et al., 2007). Conversely, inaccuracy of a measure could also lead to under-identification of students by producing false negatives. In this case, students who need extra instruction may not receive appropriate supports early on and fall further behind compared to their peers (Jenkins et al., 2007). How well a screener performs depends on its sensitivity (i.e., ability to distinguish between true positives and false positives) and its specificity (i.e., ability to distinguish between true negatives and false negatives). The sensitivity and specificity of published screens are almost always reported in their technical manuals; however, these are limited to the cut-points, sample information (i.e., base rate), and outcome measure used, making generalization difficult.

While sensitivity and specificity are widely reported indices for screeners, two lesser known but equally important indices are positive predictive power (PPP) and negative predictive power (NPP; Schatschneider et al., 2008). PPP refers to the percentage of students who will have a reading problem out of all students identified as being at risk based upon their screening performance. Conversely, NPP refers to the percentage of students who will not have a reading problem out of all students identified as being not at risk based upon the screen. Unlike sensitivity and specificity, which are thought to be insensitive to the base rate of the problem, PPP and NPP vary with the base rate. Specifically, if the base rate of reading problems is low, then PPP will be lower, meaning the screen will identify an increasingly large number of students as being at risk when in fact they are not. It is the PPP and the NPP that will inform how well a screen will perform in a particular educational context. All four of these indices (sensitivity, specificity, PPP, and NPP) are needed to evaluate how a screen will perform.

Screeners are used in a variety of educational settings to determine which students are at risk for not meeting a benchmark (such as passing an end-of-the-year assessment) and how resources will be allocated by determining which students will receive extra services. Despite the widespread use of screeners for such decisions, the classification ability of many screeners is not suited for such an important job. Jenkins and Johnson (2008) specify that a good screener in reading should correctly identify 80% or more of the students who are not at risk (specificity > 0.80) and correctly identify at least 90% of students who are at risk for reading failure (sensitivity > 0.90). However, screeners can be far from this ideal (e.g., Riedel, 2007; Schatschneider, 2006).

In addition to the poor performance of current screeners, selecting the right screener for a given school or district can be difficult. Screeners are based on assessments measured on a continuous scale with pre-identified cut-points used to define the at risk group. Obtaining classification information typically consists of forming groups of students that are below a given cut-point on the screener and assessing what percentage of those students are below the cut-point on the outcome. These comparisons form a traditional 2×2 confusability matrix which contains four rates: (a) true positive (i.e., those who are identified as having a problem and actually do); (b) true negative (i.e., those who are identified as not having a problem and don't); (c) false positive (i.e., those who are identified as having a problem but don't); and (d) false negative (i.e., those who are identified as not having a problem but do). Using these rates, numerous classification indices can be obtained such as correct classification, sensitivity, specificity, PPP, and NPP.

Many screeners report classification information from a test sample (National Center on Intensive Intervention, n.d.) with a given cut-point and base rate; however, these test samples do not always generalize well to other groups of children in a particular school or district given that the base rate in a sample greatly influences classification accuracy. The purpose of this paper is to demonstrate that if bivariate normality is assumed, it is possible to derive all of the classification information needed to evaluate a screener from simply knowing the correlations of the original variables from which the screener and outcome are derived and the cut-points chosen for group membership. Moreover, we have used this information to make a free online tool that allows users to input correlation and cut-point information to determine what the classification information associated with that information would be. Furthermore, with this approach, we demonstrate that the strength of the original correlation between the screener and outcome measure must be very high to obtain the desired classification accuracy. This helps to inform development of which screeners will yield good classification for specific outcomes as well as exemplifying the need for measures with high reliability given that the maximum correlation that can be observed is limited to the internal reliability of the lower of the two measures (Nunnally & Bernstein, 1994). For example, if the reliability of a screener is 0.7, the maximum correlation that can be observed with any variable is the square root of 0.7, or 0.84, which is insufficient for quality classification. Additionally, the free online tool allows for a user-friendly application to explore these classification properties without requiring an understanding of the underlying formulas, allowing administrators and policymakers to try out varying cut-points to determine which are ideal for their needs.

Calculation of classification indices from correlation and cut-points

Here we demonstrate that if bivariate normality is assumed, it is possible to derive all of the classification information needed to evaluate a screener from simply knowing the correlations of the original variables from which the screener and outcome are derived and the cut-points chosen for determining who is at risk for a poor outcome.

The cumulative probability distribution function for a bivariate normal distribution is given as

$$CDFBinorm = \frac{1}{2\pi\sqrt{1-r2}} \iint_{-\infty}^{xandy} exp\left[-\frac{u^2 - 2ruv + v^2}{2(1-r^2)}\right] dvdu$$

with X and Y being standard continuous normal variables with a mean of zero and a standard deviation of one, r representing the correlation between the two variables, and u and v representing random variables x and y. This formula yields the probability that an observation will be located in the region below x and y (in this case, below the cut-point on the screener and outcome).

If we assume a bivariate normal distribution between the screener and the outcome, we can calculate all of the classification indices if we know the correlation between the two measures. The true positive proportion (tpp) can be calculated using the above formula where x and y are the screen and outcome cut-points in a continuous normal distribution because it provides the probability that data will fall below these cut-points (i.e., the proportion of the total sample that falls below those cut-points). Similarly, the true negative proportion (tnp) can be calculated by multiplying x and y by -1. This provides the proportion of the total sample that is above both cut-points. The false positive proportion (fpp) can be calculated by subtracting the true positive proportion from the proportion of data that would fall below the cut-point on the screener given a normal distribution. False

negative proportion (fnp) can be calculated by subtracting the true positive proportion from the proportion of data that would fall below the cut-point on the outcome given a normal distribution. Most classification indices can be obtained by knowing just these four values, including correct classification, sensitivity, specificity, PPP, and NPP.

Correct classification can be calculated by adding the true positive proportion with the true negative proportion. This provides the proportion of the total sample that was correctly identified (agreement between the screen and outcome).

$$CorrectClassification = tpp + tnp$$

Sensitivity (also referred to as true positive rate) is calculated by dividing the true positive by the sum of the true positive and false negative. This provides the proportion of the sample below the cut-point on the outcome that is also below the cut-point on the screener. In the example of reading disability identification, this would be the proportion of individuals with a reading disability that were correctly identified as being at risk by the screener.

$$Sensitivity = \frac{tpp}{tpp + fnp}$$

Specificity (also referred to as true negative rate) is calculated by dividing the true negative by the sum of the true negative and false positive. This provides the proportion of the sample above the cut-point on the outcome that is also above the cut-point on the screener. This would be the proportion of individuals without a reading disability that were classified as not at risk for having a reading disability by the screener.

$$Specificity = \frac{tnp}{tnp + fpp}$$

Positive predictive power is calculated by dividing the true positive by the sum of the true positive and false positive. This provides the proportion of the sample below the cut-point on the screener that is also below the cut-point on the outcome. This reflects the proportion of individuals identified as at risk for reading disability by the screener that actually have a reading disability.

$$PPP = \frac{tpp}{tpp + fpp}$$

Negative predictive power is calculated by dividing the true negative by the sum of the true negative and false negative. This provides the proportion of those above the cut-point on the screener that are also above the cut-point on the outcome. This reflects the proportion of individuals classified as not at risk for having a reading disability by the screener who do not have a reading disability.

$$NPP = \frac{tnp}{tnp + fnp}$$

Using a Shiny app to calculate classification indices

Using these formulas, we created a tool which automatically calculates all classification indices for a given correlation and cut-points on continuous normally distributed screener and outcome measures. The accompanying free online tool allows users to obtain all of the

classification criteria described above given the screening needs for their specific sample. The tool allows for cut-points to be input as *z*-scores, percentile ranks, or standard scores. After users input the correlation and cut-points, all of the above classification information is presented in a table. This tool can be used by researchers, administrators, and policymakers to explore how a screener will perform for an outcome on a given sample prior to implementation on that sample simply by knowing the correlation between the measures from which the screener and outcome are derived, the base rate in the sample, and the desired cut-point on the screener. This can also inform what an ideal cut-point may be for a given correlation, base rate, and desired classification criteria. This free online tool can be found at https://qmi-fcrr.shinyapps.io/Correlations_Cut-Points_Classification/ (see Fig. 1 for a screenshot of the tool layout).

Evaluating estimation accuracy

Using a dataset of 28,389 children attending Florida public schools which has been previously used in papers evaluating screening (Brown Waesche et al., 2011; Spencer et al., 2014), we demonstrate how close the estimated values are to those observed in the empirical dataset.

Method

Dataset

The dataset contained a sample of 28,389 students attending Florida public school in 1st grade. Data from 21,523 of these students were also available in 2nd grade. Fall Dynamic Indicators of

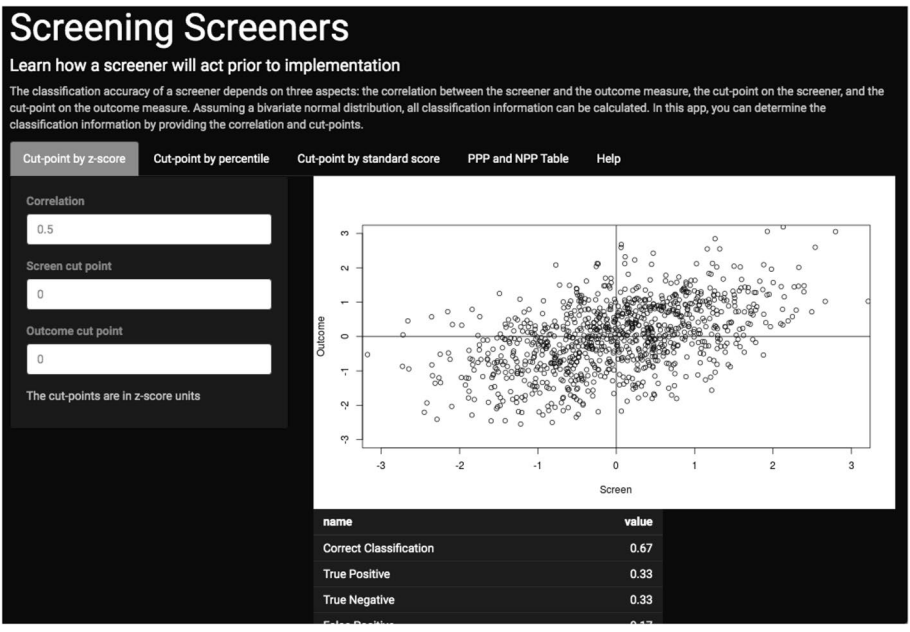

Fig. 1 Screen capture of the free online tool available at https://qmi-fcrr.shinyapps.io/Correlations_Cut-Points_Classification/

Basic Early Literacy Skills (DIBELS) measures were used to screen for risk of being in the bottom 25% on the end-of-the-year assessment (Stanford Achievement Test 10th edition (SAT10)). More information on the sample and measures can be found in Spencer et al. (2014) but are simply used as an example to demonstrate estimation accuracy of these procedures. For the purposes of this demonstration, only students with complete data for screening and outcome measures were included in the analysis. We used four possible cut-points (10, 20, 25, and 40%) to examine the estimation accuracy of these procedures, representing four common cut-points used in screening. An infinite number of combinations can be made but we present the results of 24 conditions representing possible screening scenarios in order to demonstrate how close the estimated values are to those observed in the empirical sample.

Screening measures

Letter Naming Fluency Letter Naming Fluency (LNF) is an individually administered measure from the DIBELS (Good et al., 2001) in which students are presented with a page of upper- and lowercase letters arranged in a random order and are asked to name as many as they can in one minute. LNF was assessed in the fall. The alternative-form reliability of this measure is 0.88 (Good et al., 2001). Test–retest reliability completed on a subset of 2408 children from this sample was 0.90 (Catts et al., 2009).

Phoneme Segmentation Fluency Phoneme Segmentation Fluency (PSF) is an individually administered measure of phonological awareness from DIBELS which assesses a student's ability to segment three or four phoneme words into their individual phonemes. The score consists of the number of phonemes correctly produced in one minute. PSF was assessed in the fall. The alternative-form reliability of this measure is 0.79 (Good et al., 2001). Test–retest reliability completed on a subset of 2408 children from this sample was 0.69 (Catts et al., 2009).

Nonsense Word Fluency DIBELS Nonsense Word Fluency (NWF) is an individually administered test of the alphabetic principle including letter-sound correspondence and the ability to blend letters into words with which letters are being represented by their common sounds (Kaminski & Good, 1996). Students are presented with a paper containing randomly ordered nonsense words and asked to produce the individual letter sounds of each letter verbally or read the whole nonsense word allow. The score represents the number of letter-sounds produced correctly in one minute. NWF was assessed in the fall. The alternative-form reliability of this measure is 0.83 (Good et al., 2001). Test–retest reliability completed on a subset of 2408 children from this sample was 0.86 (Catts et al., 2009).

Oral Reading Fluency DIBELS Oral Reading Fluency (ORF) is an individually administered measure of accuracy and fluency of reading connected text in which the score represents the number of correct words read from the passage in a minute. ORF was assessed in the fall. The median alternate-form reliability for oral reading of passages is 0.94 (Good et al., 2001). Test–retest reliability completed on a subset of 2408 children from this sample was 0.96 (Catts et al., 2009).

Outcome measure

Stanford Achievement Test 10th edition SAT10 (Brace, 2003) is a group administered untimed assessment of reading comprehension in which students read passages and answer

multiple-choice questions. SAT10 was assessed in the spring. The criterion cut used here was at the 25th percentile. The reliability of this measure is 0.88 (Catts et al., 2009).

Calculating observed classification criteria

Classification indices were calculated by creating dichotomous variables in which each student was either at or below a given cut-point or above that cut-point. The observed percentage of students that were at or below the cut-point can be found in the outcome cut-point column of the table, with the a priori cut-point in parentheses. For example, in grade 1, 27% of the students were at or below the a priori cutoff of the 25th percentile.

Calculating estimated classification criteria using observed sample cut-points

Using the observed correlation between the screening and outcome measures and the percent of the sample below the cut-point on each, classification indices were calculated using the estimation procedures described above yielding values that would be obtained using the Shiny app.

Calculating estimated classification criteria using a priori cut-points

Using the observed correlation between the screening and outcome measures and the a priori cut-point (10, 20, 25, 40%), classification indices were calculated using the estimation procedures described above yielding values that would be obtained using the Shiny app.

Results

Observed and estimated classification indices are presented in Table 1. The estimated values using the sample cut-points were very close to the observed values, with absolute errors ranging from 0.004 to 0.109 with a mean of 0.021. These results demonstrate that it is possible to closely estimate classification criteria for a screener given the correlation between screening and outcome measure and the cut-point on each. Using the a priori cutoffs (demonstrating an example of when one may not have sample distribution information ahead of time), the absolute errors between estimate and observed values ranged from 0.001 to 0.157 with a mean of 0.034.

The difference between the a priori cut-point and the observed percent of the sample at or below the cut-point is due to what is referred to as measure coarseness. That is, a large portion of children received the score at the cut-point. An excellent example of this is the ORF measure in which 16% of the sample received a score of 0. Furthermore, this case exemplifies how non-normality can impact the prediction. The 1st grade ORF screening measure had a large skew (as seen in Fig. 2), leading to an inability to differentiate children at the bottom of the distribution. Despite this, the estimated classification indices using the sample cut-points were still close to the observed values. However, when using the a priori cut-points to make the estimates, the values move further away because, although the intent was to estimate a cut-point of 10%, in reality the cut-point was at 16%. This is caused by the measure coarseness (16% of the children having the same score) and not specifically the non-normality. Even with the skewed distribution, when using the sample cut-points, the predicted values were quite close to the observed values, suggesting that measure coarseness may have more practical implications for estimates than non-normality.

Table 1 Results of observed values compared to predicted values

Grade	Screen	cor	Screen cut-point	Outcome cut-point	Sensitivity		Specificity		PPP		NPP		CC	
					obs	Predicted	obs	Predicted	obs	Predicted	obs	Predicted	obs	Predicted
1	LNF	0.55	42.35 (40)	27.02 (25)	0.74	0.71 (0.69)	0.69	0.68 (0.70)	0.47	0.45 (0.43)	0.88	0.86 (0.87)	0.71	0.69 (0.70)
1	LNF	0.55	26.69 (25)	27.02 (25)	0.56	0.52 (0.51)	0.84	0.83 (0.84)	0.57	0.53 (0.51)	0.84	0.82 (0.84)	0.77	0.74 (0.75)
1	LNF	0.55	21.24 (20)	27.02 (25)	0.47	0.44 (0.43)	0.88	0.87 (0.88)	0.60	0.56 (0.54)	0.82	0.81 (0.82)	0.77	0.76 (0.77)
1	LNF	0.55	10.03 (10)	27.02 (25)	0.27	0.24 (0.25)	0.96	0.95 (0.95)	0.72	0.66 (0.63)	0.78	0.77 (0.79)	0.77	0.76 (0.78)
1	PSF	0.40	40.00 (40)	27.02 (25)	0.64	0.60 (0.60)	0.69	0.67 (0.67)	0.43	0.40 (0.38)	0.84	0.82 (0.84)	0.68	0.65 (0.65)
1	PSF	0.40	25.2 (25)	27.02 (25)	0.48	0.42 (0.43)	0.83	0.81 (0.81)	0.51	0.45 (0.43)	0.81	0.79 (0.81)	0.74	0.71 (0.71)
1	PSF	0.40	20.27 (20)	27.02 (25)	0.41	0.35 (0.36)	0.87	0.85 (0.85)	0.55	0.47 (0.45)	0.80	0.78 (0.80)	0.75	0.72 (0.73)
1	PSF	0.40	10.96 (10)	27.02 (25)	0.26	0.21 (0.20)	0.95	0.93 (0.93)	0.64	0.53 (0.51)	0.77	0.76 (0.78)	0.76	0.74 (0.75)
1	NWF	0.60	41.33 (40)	27.02 (25)	0.77	0.73 (0.73)	0.72	0.70 (0.71)	0.50	0.48 (0.45)	0.89	0.88 (0.89)	0.73	0.71 (0.71)
1	NWF	0.60	26.58 (25)	27.02 (25)	0.59	0.55 (0.54)	0.85	0.84 (0.85)	0.60	0.56 (0.54)	0.85	0.83 (0.85)	0.78	0.76 (0.77)
1	NWF	0.60	20.9 (20)	27.02 (25)	0.50	0.46 (0.46)	0.90	0.89 (0.89)	0.65	0.60 (0.58)	0.83	0.82 (0.83)	0.79	0.77 (0.78)
1	NWF	0.60	10.47 (10)	27.02 (25)	0.29	0.27 (0.27)	0.96	0.96 (0.96)	0.74	0.70 (0.68)	0.78	0.78 (0.80)	0.78	0.77 (0.79)
1	ORF	0.58	40.05 (40)	27.02 (25)	0.80	0.70 (0.71)	0.75	0.71 (0.70)	0.54	0.47 (0.45)	0.91	0.87 (0.88)	0.76	0.71 (0.71)
1	ORF	0.58	25.98 (25)	27.02 (25)	0.61	0.53 (0.53)	0.87	0.84 (0.84)	0.63	0.55 (0.53)	0.86	0.83 (0.84)	0.80	0.76 (0.76)
1	ORF	0.58	21.24 (20)	27.02 (25)	0.52	0.46 (0.45)	0.90	0.88 (0.88)	0.67	0.58 (0.56)	0.84	0.81 (0.83)	0.80	0.76 (0.78)
1	ORF	0.58	16.34 (10)	27.02 (25)	0.42	0.38 (0.26)	0.93	0.92 (0.95)	0.69	0.62 (0.66)	0.81	0.80 (0.80)	0.79	0.77 (0.78)
2	NWF	0.50	40.43 (40)	26.08 (25)	0.69	0.66 (0.66)	0.70	0.69 (0.69)	0.45	0.43 (0.41)	0.87	0.85 (0.86)	0.70	0.68 (0.68)
2	NWF	0.50	26.36 (25)	26.08 (25)	0.53	0.49 (0.48)	0.83	0.82 (0.83)	0.53	0.49 (0.48)	0.83	0.82 (0.83)	0.75	0.73 (0.74)
2	NWF	0.50	21.1 (20)	26.08 (25)	0.46	0.42 (0.41)	0.88	0.86 (0.87)	0.57	0.52 (0.51)	0.82	0.81 (0.81)	0.77	0.75 (0.75)
2	NWF	0.50	10.84 (10)	26.08 (25)	0.27	0.25 (0.23)	0.95	0.94 (0.94)	0.66	0.59 (0.59)	0.79	0.78 (0.79)	0.77	0.76 (0.77)
2	ORF	0.67	40.13 (40)	26.08 (25)	0.81	0.76 (0.77)	0.74	0.73 (0.72)	0.53	0.49 (0.48)	0.92	0.90 (0.90)	0.76	0.74 (0.73)
2	ORF	0.67	25.38 (25)	26.08 (25)	0.65	0.58 (0.58)	0.89	0.86 (0.86)	0.67	0.59 (0.58)	0.88	0.85 (0.86)	0.82	0.79 (0.79)
2	ORF	0.67	20.92 (20)	26.08 (25)	0.58	0.51 (0.50)	0.92	0.90 (0.90)	0.72	0.63 (0.62)	0.86	0.84 (0.84)	0.83	0.79 (0.80)
2	ORF	0.67	10.57 (10)	26.08 (25)	0.34	0.30 (0.29)	0.98	0.96 (0.96)	0.85	0.74 (0.73)	0.81	0.80 (0.80)	0.81	0.79 (0.80)

Values in the parentheses represent the a priori values. Values in the predicted column not in the parentheses represent the value calculated using the estimation procedures using the cut-points observed in the sample (percent of students at or below the cut-point in the sample). Values in the predicted column inside the parentheses are those that are estimated using the a priori cut-points (i.e., 10, 20, 25, or 40)

cor, correlation between screen and outcome measures; *obs*, observed value; *PPP*, positive predictive power; *NPP*, negative predictive power; *CC*, correct classification

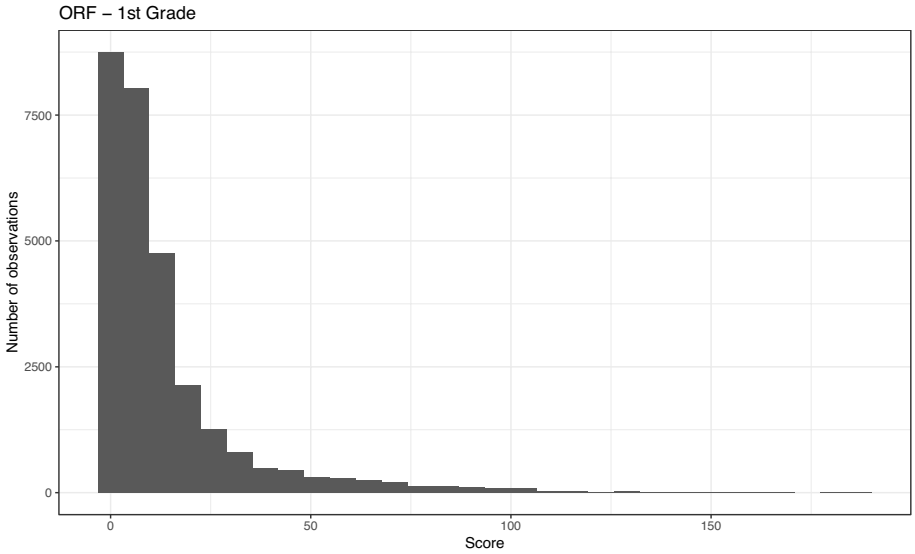

Fig. 2 Histogram of the number of students with a given score on 1st grade Oral Reading Fluency (ORF)

Evaluating the impacts of correlation and cut-points on classification indices

Method

Classification indices were calculated for 72 possible conditions (these are just a small subsample of the possible conditions but provide insight into the general relations) with correlations of 0.50, 0.60, 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95 and cut-points at the 10th, 20th, and 25th percentile. We have created tables that show the expected classification indices that would be obtained based on the combinations of these correlations and cut-points (Tables 2, 3, and 4). We then evaluated how these 72 conditions fit within the criteria for good screening set by Jenkins and Johnson (2008).

Results

Results showed that as the correlation increased so too did classification accuracy, which was expected. However, what was unexpected is the magnitude of the correlation needed for adequate classification. Of the 72 conditions presented in the tables, only six of them would meet the criteria for good classification as set by Jenkins and Johnson (2008) of having a specificity greater than 0.8 and sensitivity greater than 0.9. Of those six conditions, three had a correlation of 0.95, two had a correlation of 0.90, and one had a correlation of 0.85. This demonstrates the need for very high correlations between a screener and outcome to achieve good screening accuracy. Additionally, the extent to which the correlation impacts classification accuracy should be noted. Take a correlation of 0.6 for example, which is typically regarded to be a moderate correlation. With a correlation of 0.6 and cut-points at the 20th percentile, half of the children that are identified as at risk are actually not at risk. In an educational setting,

Table 2 Results with base rate of 10% at varying correlations and screening cut-points

Correlation	Screen cut-point	Outcome cut-point	Correct classification	Sensitivity	Specificity	Positive predictive power	Negative predictive power
0.5	10	10	0.86	0.32	0.92	0.32	0.92
0.5	20	10	0.80	0.51	0.83	0.26	0.94
0.5	25	10	0.77	0.59	0.79	0.24	0.95
0.6	10	10	0.88	0.39	0.93	0.39	0.93
0.6	20	10	0.82	0.60	0.84	0.30	0.95
0.6	25	10	0.78	0.67	0.80	0.27	0.96
0.7	10	10	0.89	0.47	0.94	0.47	0.94
0.7	20	10	0.84	0.69	0.85	0.34	0.96
0.7	25	10	0.80	0.76	0.81	0.31	0.97
0.75	10	10	0.90	0.51	0.95	0.51	0.95
0.75	20	10	0.85	0.74	0.86	0.37	0.97
0.75	25	10	0.81	0.81	0.81	0.32	0.97
0.8	10	10	0.91	0.56	0.95	0.56	0.95
0.8	20	10	0.86	0.79	0.87	0.40	0.97
0.8	25	10	0.82	0.86	0.82	0.34	0.98
0.85	10	10	0.92	0.62	0.96	0.62	0.96
0.85	20	10	0.87	0.85	0.87	0.43	0.98
0.85	25	10	0.83	0.91	0.82	0.36	0.99
0.9	10	10	0.94	0.69	0.97	0.69	0.97
0.9	20	10	0.88	0.91	0.88	0.46	0.99
0.9	25	10	0.84	0.96	0.83	0.38	0.99
0.95	10	10	0.96	0.78	0.98	0.78	0.98
0.95	20	10	0.89	0.97	0.89	0.49	1.00
0.95	25	10	0.85	0.99	0.83	0.40	1.00

this would mean providing double the amount of extra services than are actually needed. Furthermore, the false negative rate is 0.5, meaning that half of the children that have reading difficulties are not identified as such, i.e., that half of the children that need services would not receive them. However, as the correlation increases, so does the classification accuracy. If we use the same cut-points (20th%ile) and increase the correlation from 0.6 to 0.9, we increase the number of those correctly identified as having a problem from 50 to 75% and the number of those correctly identified as not having a problem from 50 to 94%. Even at a correlation of 0.9, we are still only accurately classifying 90% of the children.

These six conditions with a sensitivity above 0.9 and specificity above 0.8 also demonstrate another feature important to classification: base rate. The one condition with a correlation of 0.85 that maintained good screening had a base rate (outcome cut-point) at the 10th percentile and a screening cut-point at the 25th percentile. Although the sensitivity is 0.91 and specificity is 0.82, meeting the criteria set by Jenkins and Johnson (2008), the positive predictive power is 0.36, meaning that only 36% of the children classified as at risk actually are. Contextualizing this in cases when screening is used to identify students needing extra services, for 100 students identified as needing extra services, 64 of those students did not. This may be especially impactful when resources are low and there are not enough resources for all 100 students in order to

Table 3 Results with base rate of 20% at varying correlations and screening cut-points

Correlation	Screen cut-point	Outcome cut-point	Correct classification	Sensitivity	Specificity	Positive predictive power	Negative predictive power
0.5	10	20	0.80	0.26	0.94	0.51	0.83
0.5	20	20	0.77	0.44	0.86	0.44	0.86
0.5	25	20	0.75	0.51	0.81	0.41	0.87
0.6	10	20	0.82	0.30	0.95	0.60	0.84
0.6	20	20	0.80	0.50	0.87	0.50	0.87
0.6	25	20	0.78	0.57	0.83	0.46	0.89
0.7	10	20	0.84	0.34	0.96	0.69	0.85
0.7	20	20	0.83	0.56	0.89	0.56	0.89
0.7	25	20	0.81	0.65	0.85	0.52	0.91
0.75	10	20	0.85	0.37	0.97	0.74	0.86
0.75	20	20	0.84	0.60	0.90	0.60	0.90
0.75	25	20	0.83	0.69	0.86	0.55	0.92
0.8	10	20	0.86	0.40	0.97	0.79	0.87
0.8	20	20	0.86	0.65	0.91	0.65	0.91
0.8	25	20	0.84	0.73	0.87	0.59	0.93
0.85	10	20	0.87	0.43	0.98	0.85	0.87
0.85	20	20	0.88	0.69	0.92	0.69	0.92
0.85	25	20	0.86	0.78	0.88	0.63	0.94
0.9	10	20	0.88	0.46	0.99	0.91	0.88
0.9	20	20	0.90	0.75	0.94	0.75	0.94
0.9	25	20	0.89	0.84	0.90	0.67	0.96
0.95	10	20	0.89	0.49	1.00	0.97	0.89
0.95	20	20	0.93	0.82	0.96	0.82	0.96
0.95	25	20	0.91	0.91	0.92	0.73	0.98

ensure that those 36 are receiving services. When the base rate in a sample is lower, there is an increased likelihood for false positives, making base rate an important consideration when evaluating a screener. When a smaller proportion of individuals in the sample have the identifying characteristic, it becomes increasingly difficult for the screener to correctly distinguish those individuals from the rest. Returning to our example of a correlation of 0.6 with cut-points at the 20th percentile, the positive predictive power is 0.5, meaning that half of the time that a person is identified by a screener as being at risk for reading difficulties, they will not actually develop reading difficulties. If we lower the base rate (and screening cut-point) to the 10th percentile, assuming the same 0.6 correlation, the positive predictive power lowers to 0.3, meaning that only a third of those identified as needing extra services actually do. Thus, base rate in the sample is important to take into consideration when selecting a screener. When the base rate is low, an increase in false positives is observed. One way to handle this is the implementation of gated screening (Meehl & Rosen, 1955; Van Norman et al., 2019). Within gated screening, an initial screen is conducted that eliminates children that are not likely to experience reading difficulties. The rest however, rather than being labeled or assigned additional services, undergo a second screening process with a more extensive battery. Since a number of negative children are removed from the sample from the first screen, the base rate in the sample is higher, decreasing

Table 4 Results with base rate of 25% at varying correlations and screening cut-points

Correlation	Screen cut-point	Outcome cut-point	Correct classification	Sensitivity	Specificity	Positive predictive power	Negative predictive power
0.5	10	25	0.77	0.24	0.95	0.59	0.79
0.5	20	25	0.75	0.41	0.87	0.51	0.81
0.5	25	25	0.74	0.48	0.83	0.48	0.83
0.6	10	25	0.78	0.27	0.96	0.67	0.80
0.6	20	25	0.78	0.46	0.89	0.57	0.83
0.6	25	25	0.77	0.54	0.85	0.54	0.85
0.7	10	25	0.80	0.31	0.97	0.76	0.81
0.7	20	25	0.81	0.52	0.91	0.65	0.85
0.7	25	25	0.80	0.60	0.87	0.60	0.87
0.75	10	25	0.81	0.32	0.97	0.81	0.81
0.75	20	25	0.83	0.55	0.92	0.69	0.86
0.75	25	25	0.82	0.64	0.88	0.64	0.88
0.8	10	25	0.82	0.34	0.98	0.86	0.82
0.8	20	25	0.84	0.59	0.93	0.73	0.87
0.8	25	25	0.84	0.68	0.89	0.68	0.89
0.85	10	25	0.83	0.36	0.99	0.91	0.82
0.85	20	25	0.86	0.63	0.94	0.78	0.88
0.85	25	25	0.86	0.72	0.91	0.72	0.91
0.9	10	25	0.84	0.38	0.99	0.96	0.83
0.9	20	25	0.89	0.67	0.96	0.84	0.90
0.9	25	25	0.89	0.77	0.92	0.77	0.92
0.95	10	25	0.85	0.40	1.00	0.99	0.83
0.95	20	25	0.91	0.73	0.98	0.91	0.92
0.95	25	25	0.92	0.84	0.95	0.84	0.95

the likelihood of a false positive. Of note, while the correlation between screening and outcome measures affects classification accuracy in both single-timepoint and gated screening systems, the latter is also impacted by the correlation between the screening measures selected for each gate. For example, van Norman et al. (2019) found that when the correlation between each screening measure and the outcome was 0.90, the rate of false positives increased as the correlation between the screening measure at gate 1 and the screening measure at gate 2 approached 1. Thus, a gated screening approach aimed at reducing false positives may be most effective when the measures at each gate are highly predictive of the outcome measure but explain unique portions of its variance.

Discussion

In this paper, we have demonstrated that if bivariate normality is assumed, it is possible to derive all of the classification information needed to evaluate a screener from simply knowing the correlations of the original variables that the screener and outcome are derived

from and the cut-point on each. Using this information, we demonstrated the importance of a very high correlation between the screen and the outcome, higher than what is often observed in practice. In a meta-analysis, January and Klingbeil (2020) evaluated the overall classification accuracy of six types of early reading screeners (i.e., screeners that assessed fluency in first sounds, letter names, letter sounds, phoneme segmenting, word identification, and nonsense words) and found correlations between screeners and outcomes ranged between $r=0.35$ and $r=0.831$, with phoneme segmentation yielding the lowest and word identification the highest correlation.

In order for correlations of such magnitude to be observed in practice, it is not only important to take into account the construct of interest and its relation to the screener, but also the reliability of each of the measures involved. Given the high correlation needed for good classification accuracy, the need for measures with high reliability is emphasized because the maximum correlation that can be observed is limited by the internal reliability of the lower of the two measures. This means that if the reliability of one of the measures is 0.7, then the maximum correlation that can be observed between the two is the square root of 0.7 (0.84). Thus, the reliabilities of both measures need to be high (higher than what is currently observed in practice) if such high correlations are to be obtained. Additionally, just as increasing the number of items in a measure can help increase reliability, increasing the number of measures in a screener can also help increase reliability. Thus, the use of multiple measures with reasonable reliabilities can improve the classification accuracy of a screener. However, this still requires a high correlation between the outcome and the screening measures.

Limitations

Given space constraints, we were only able to report results of 72 scenarios varying the correlation between screener and outcome and the cut-point on each. There are an infinite number of combinations that can be made. We encourage readers to use the Shiny app to explore additional cases.

We only used a single base rate and outcome measure and limited cut-points to validate the estimation procedures. Additional work should investigate the estimation accuracy under other conditions.

Implications for practice

Many states have laws in place that encourage early identification of students with or at risk for dyslexia, with the intention to provide appropriate intervention early. Education administrators at all levels (i.e., school, district, and state) depend on screening measures for this identification. Having inaccurate measures could either lead to an over-expenditure of available resources due to over-identification (Jenkins et al., 2007), or to many children not receiving their needed intervention support due to under-identification. Classification information based on a screener's normative sample does not easily generalize to a school or district's specific information. Therefore, administrators and policymakers will benefit from knowing classification information is influenced by the properties of the relation between a screening measure and an outcome. The online tool allows users to explore the impact that various cut-points on a screener may have and select those that will benefit the majority of their students. By forming guidelines more in line with local situations, more

students with or at risk for dyslexia can be identified early and provided with appropriate interventions to mitigate its impact.

Choosing a practical screening mechanism to identify students at risk for reading disabilities is a complex issue involving decisions related to the allocation of resources. One critical component for all screeners is their classification accuracy. As we have shown, the correlation between the screener and the outcome measure is paramount. Equally important is the base rate in a specific sample: a low base rate will likely lead to an over-identification of at risk students. In this case, gated screening may be preferred (Meehl & Rosen, 1955; Van Norman et al., 2019).

Funding This research was supported by Grant P50 HD052120 for the Eunice Kennedy Shriver National Institutes of Child Health and Human Development and Grant R305B200020 from the Institute of Education Sciences.

Declarations

Conflict of interest The authors declare no competing interests.

References

Brace, H. (2003). *Stanford achievement test—Tenth edition: Technical data report*. Author.

Brown Waesche, J. S., Schatschneider, C., Maner, J. K., Ahmed, Y., & Wagner, R. K. (2011). Examining agreement and longitudinal stability among traditional and RTI-based definitions of reading disability using the affected-status agreement statistic. *Journal of Learning Disabilities, 44*(3), 296–307.

Catts, H. W., & Petscher, Y. (2018). Early identification of dyslexia current advancements and future directions. *Perspectives on Language and Literacy*, *44*(3).

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implications. *Language, Speech, and Hearing Services in Schools, 32*, 38–50.

Catts, H. W., Petscher, Y., Schatschneider, C., Sittner Bridges, M., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*(2), 163–176.

Gearin, B., Turtura, J., Kame'enui, E. J., Nelson, N. J., & Fien, H. (2018). A multiple streams analysis of recent changes to state-level dyslexia education law. *Educational Policy*. https://doi.org/10.1177/0895904818807328

Good, R. H., Kaminski, R. A., Smith, S., Laimon, D., & Dill, S. (2001). *Dynamic indicators of basic early literacy skills* (5th ed.). University of Oregon.

Individuals with Disabilities Education Act, 20 U.S.C § 1400 et seq (2012). https://uscode.house.gov/view.xhtml?path=/prelim@title20/chapter33&edition=prelim. Accessed Jan 2022

January, S.-A.A., & Klingbeil, D. A. (2020). Universal screening in grades K-2: A systematic review and meta-analysis of early reading curriculum-based measures. *Journal of School Psychology, 82*, 103–122. https://doi.org/10.1016/j.jsp.2020.08.007

Jenkins, J. R., & Johnson, E. (2008). Universal screening for reading problems: Why and how should we do this. *RTI Action Network*.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a Response to Intervention framework. *School Psychology Review, 36*(4), 582–600. https://doi.org/10.1080/02796015.2007.12087919

Kaminski, R. A., & Good, R. H., III. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25*(2), 215–227.

Kent, S. C., Wanzek, J., & Yun, J. (2019). Screening in the upper elementary grades: Identifying fourth-grade students at-risk for failing the state reading assessment. *Assessment for Effective Intervention, 44*(3), 160–172. https://doi.org/10.1177/1534508418758371

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, pattern, and cutting scores. *Psychological Bulletin, 52*, 194–216.

National Center on Improving Literacy. (n.d.). *State of dyslexia*. https://improvingliteracy.org/state-of-dyslexia. Accessed Jan 2022

National Center on Intensive Intervention. (n.d.). *Academic Screening Tools Chart* https://charts.intensiveintervention.org/ascreening. Accessed Jan 2022

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometic theory (3rd)*. McGraw-Hill.

Petscher, Y., Truckenmiller, A., & Zhou, C. (2016). *The earlier assessment for reading success [web application software]*. Florida State University.

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*(4), 546–567. https://doi.org/10.1598/RRQ.42.4.5

Schatschneider, C. (2006). *Reading difficulties: Classification and issues of prediction.* [conference paper]. Pacific Coast Research Conference, San Diego, CA.

Schatschneider, C., Petscher, Y., & Williams, K. M. (2008). How to evaluate a screening process: The vocabulary of screening and what educators need to know. In L. M. Justice & C. Vukelich (Eds.), Achieving excellence in preschool literacy instruction (pp. 304–316). The Guilford Press

Spencer, M., Wagner, R. K., Schatschneider, C., Quinn, J. M., Lopez, D., & Petscher, Y. (2014). Incorporating RTI in a hybrid model of reading disability. *Learning Disability Quarterly, 37*(3), 161–171.

Stevenson, N. A. (2017). Comparing curriculum-based measures and extant datasets for universal screening in middle school reading. *Assessment for Effective Intervention, 42*(4), 195–208. https://doi.org/10.1177/1534508417690399

Thomas, A. S., & January, S.-A.A. (2021). Evaluating the criterion validity and classification accuracy of universal screening measures in reading. *Assessment for Effective Intervention*. https://doi.org/10.1177/1534508419857232

Van Norman, E. R., Nelson, P. M., Klingbeil, D. A., et al. (2019). Gated screening frameworks for academic concerns: The influence of redundant information on diagnostic accuracy outcomes. *Contemporary School Psychology, 23*, 152–162. https://doi.org/10.1007/s40688-018-0183-0

Youman, M., & Mather, N. (2018). Dyslexia laws in the USA: A 2018 update. *Perspectives on Language and Literacy, 44*(2), 37–41.