Title
Are two screeners better than one? A simulation study of correlation and classification in universal screening

Citation:
White, C. M. & Schatschneider, C. (2023). Are two screeners better than one? A simulation study of correlation and classification in universal screening. *Contemporary School Psychology*. Online first publication. https://doi.org/10.1007/s40688-023-00478-0

Authors
Christine M. White[1,2] (ORCID: 0000-0002-8913-282)
Christopher Schatschneider[1,2] (ORCID: 0000-0002-1700-7685)

Author Affiliations
1. Florida State University, Department of Psychology, Tallahassee, FL, USA.
2. Florida Center for Reading Research, Tallahassee, FL, USA

Corresponding Author
Christine M. White. Address: Department of Psychology, 1107 West Call Street, Tallahassee, FL, 32306. Email Address: white@psy.fsu.edu

Publication Date
Journal: Contemporary School Psychology
Online first publication date: 03 October 2023

Statements and Declarations

Competing Interests
On behalf of all authors, the corresponding author states that there is no conflict of interest. The authors have no relevant financial or non-financial interests to disclose.

Funding
The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200020 to the Florida Center for Reading Research at Florida State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Acknowledgments
The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200020 to the Florida Center for Reading Research at Florida State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Author Contributions
Both authors contributed to the study conception and design. Simulations were conducted by Christine M. White with advising from Christopher Schatschneider. The first draft of the manuscript was written by Christine M. White and Christopher Schatschneider commented on previous versions of the manuscript. Both authors read and approved the final manuscript.

Ethical Approval, Consent to Participate, Consent to Publish
The research reported here used simulations to generate data. The above disclosures do not apply.

Code Availability Statement. The R code and associated output supporting the findings of the study have been made openly available on the Open Science Framework at https://osf.io/xwyr8/.

Abstract

Universal screening to predict students' risk for reading problems is a foundational component of the Multi-Tiered Systems of Support framework and is required by law in many U.S. states. School or district administrators are tasked with selecting screening assessments that are both technically adequate and feasible given the resources of their local context. One common recommendation is that educational screening assessments should have at least a sensitivity of .9 and a specificity of .8. The two studies presented here used simulation methodology to identify the screener-outcome correlation(s) needed to achieve these recommended levels of sensitivity and sensitivity with a one-indicator (Study 1) or two-indicator (Study 2) screening battery. In both studies, the base rates of non-proficiency were manipulated. Results showed that the minimum correlations needed to achieve this recommendation were higher than what is typically observed in practice, and also varied across samples with differing base rates. Furthermore, screening assessments with the recommended levels of sensitivity and specificity had high rates of false positive classifications that depended on the base rate, cut-point, and method of assigning risk. These results suggest that the practice of issuing specific criteria for the sensitivity and specificity of screeners may be misguided. Implications for the evaluation of the technical adequacy of screening assessments and recommendations for practitioners are discussed.

Keywords: universal screening - classification - simulation - education

Title: Are two screeners better than one? A simulation study of correlation and classification in universal screening

Abstract

Universal screening to predict students' risk for reading problems is a foundational component of the Multi-Tiered Systems of Support framework and is required by law in many U.S. states. School or district administrators are tasked with selecting screening assessments that are both technically adequate and feasible given the resources of their local context. One common recommendation is that educational screening assessments should have at least a sensitivity of .9 and a specificity of .8. The two studies presented here used simulation methodology to identify the screener-outcome correlation(s) needed to achieve these recommended levels of sensitivity and sensitivity with a one-indicator (Study 1) or two-indicator (Study 2) screening battery. In both studies, the base rates of non-proficiency were manipulated. Results showed that the minimum correlations needed to achieve this recommendation were higher than what is typically observed in practice, and also varied across samples with differing base rates. Furthermore, screening assessments with the recommended levels of sensitivity and specificity had high rates of false positive classifications that depended on the base rate, cut-point, and method of assigning risk. These results suggest that the practice of issuing specific criteria for the sensitivity and specificity of screeners may be misguided. Implications for the evaluation of the technical adequacy of screening assessments and recommendations for practitioners are discussed.

Keywords: universal screening - classification - simulation - education

Introduction

Many students in the United States do not attain proficiency in reading, with nearly 35% of 4th graders performing below a "Basic" level of reading comprehension and 65% reading below grade-level on a recent national assessment (NCES, 2019). This represents a public health issue as reading ability affects not only academic but societal and health outcomes (Sanfilippo et al., 2020). Waiting until students have begun to experience reading difficulties or until a formal dyslexia diagnosis is conferred (typically not until the end of 2nd grade) can have harmful consequences, as falling behind one's peers in reading can lead to academic, social, and emotional problems which pervade and compound over time. Further, reading interventions are most effective at improving outcomes when delivered in kindergarten or first grade (Ozernov-Palichik & Gaab, 2016). Thus, over the past few decades, empirical research has focused on the estimation of risk for reading problems before the onset of reading instruction and movement toward a prevention model of dyslexia (Catts et al., 2013; Catts & Hogan, 2020).

Universal screening is an essential part of achieving early identification or prevention of reading difficulties at scale. Screening typically involves assessing the reading and reading-related skills of all students in an educational context to predict who is at risk of non-proficient performance on a later reading assessment so that supplemental instruction or intervention can be provided (Gaab & Petscher, 2022). In the wake of legislation such as the No Child Left Behind Act (NCLB, 2002) reauthorized as the Every Student Succeeds Act (ESSA, 2015), and the Individuals with Disabilities Education Act (IDEA, 2012), and increased adoption of Response-to-Intervention/Multi-Tiered System of Support frameworks in schools (of which universal screening is a foundational element; Troester et al., 2022; Fuchs & Vaughn, 2012), universal screening for reading problems has become common in the U.S., with over 40 states having

some provision for screening in their laws in 2023 (National Center on Improving Literacy, n.d.).

Despite the ubiquity of screening, implementation varies across educational settings (Mellard et al., 2009). This is partly due to a lack of consensus across states on the definition of dyslexia or reading problems, which literacy skills and age groups should be targeted by screening, and how often screening should occur (Gearin et al., 2021). Additionally, the large number of commercially available screening assessments drives further variability (National Center for Intensive Intervention, n.d.; Jenkins et al., 2013). In many states, the Department of Education provides a list of approved screening assessments from which districts or schools must choose; however, this usually comprises five or more options which vary in terms of cost, administration time, and technical adequacy. For example, Missouri's Department of Education lists 13 state-approved screening assessments for use in kindergarten (MDESE, n.d.). The Missouri list also demonstrates that state laws may specify which literacy skill components (e.g., phonological awareness, rapid naming, and sound symbol correspondence) are required to be targeted by screening, and then allow practitioners mix and match assessments from different publishers to satisfy these requirements. Therefore, practitioners at the district or school level still need to select the best assessment(s) for their academic setting.

Evaluation of Screening Assessments

While, in reality, many factors influence the selection of a screening assessment, education practitioners are urged to evaluate the technical adequacy of potential screeners. Technical adequacy encompasses multiple properties of a screening assessment such as reliability, validity (e.g., the correlation between a screener and the outcome it is intended to predict), and classification accuracy (Petscher & Suhr, 2022). Classification accuracy reflects

how well a screening assessment sorts students into categories: True Positives, False Positives, True Negatives, and False Negatives. These categories arise from the fact that, although reading and related abilities are usually measured on a continuum, the goal of screening is to determine whether each student should be provided with extra testing, resources, or intervention—a binary decision. As such, risk status is coded as a binary or ordinal variable, with categories defined by imposing cut-points on the distribution of scores. Traditionally, outcomes are also operationalized as binary categories: for example, schools may screen to predict which students are at risk for failing a year-end state assessment (Kent et al., 2019). In a hypothetical scenario where no intervention is provided, a screener with perfect classification accuracy would result in every student identified as "at-risk" by the screener failing the outcome (True Positives; TP), and every student identified as "not at-risk" passing the outcome (True Negatives; TN). In reality, due to measurement error and the fact that reading is an actively developing, latent ability, no screening measure can be perfectly accurate: there is always a trade-off between the rate of False Positives (FP) and False Negatives (FN; Fletcher et al., 2002).

These four numbers, sometimes in combination with the base rate of problems in a particular screening context, can be used to calculate many indices giving unique insights into how a screener is performing, such as sensitivity, specificity, likelihood ratios, positive and negative predictive power, Cohen's kappa, pre- and post-test probabilities, and so on (Streiner, 2003). However, sensitivity, specificity, and predictive power are the most intuitive metrics and those that are most commonly reported by test publishers (NCII Screening Tools Chart, n.d.). In the context of screening for risk of reading difficulties, sensitivity is the likelihood that a student who is truly at risk for a reading problem will be flagged by the screener (TP/(TP+FN)), and specificity is the likelihood that a student who truly does not have a reading problem will not be

identified by the screener (TN/(TN+FP)). Positive predictive power (PPP) is the likelihood that a student flagged by the screener is truly at-risk for a reading problem given the base rate of the condition in the sample (TP/(TP+FP)), while negative predictive power (NPP) reflects the likelihood that a student who is not flagged by the screener is truly not at-risk for a reading problem given the base rate (TN/(TN+FN)).

Sensitivity and specificity are often juxtaposed with PPP and NPP, with the former characterized as static properties of the screening assessment and the latter characterized as being sample-dependent. For example, Petscher et al. (2011) write that "if a screener was used in two separate samples where one was higher achieving than the other, similar estimates of sensitivity and specificity could be obtained, while different values for the positive and negative predictive power would be calculated" (p. 160). Thus, PPP and NPP indices are likely more useful than sensitivity and specificity for practitioners who are interested in rates of False Positive and False Negative classifications in their local context.

Most contemporary education researchers agree that the base rate of problems in the local context needs to be considered in screening and suggest that practitioners consider indices such as predictive power or post-test probabilities (which combine sensitivity, specificity, and base rate information) when evaluating screening assessments (Swets, 1992; vanDerHeyden, 2011; Klingbeil et al., 2017). However, in many cases, sensitivity and/or specificity are still emphasized (e.g., Burns et al., 2022). Further, the characterization of sensitivity and specificity as being invariant properties of tests has led researchers to issue specific recommendations for these indices, against which practitioners are ostensibly meant to compare potential screening assessments. One common suggestion is that educational screening assessments should have a sensitivity above .9 and a specificity above .8 (Jenkins & Johnson, 2008; Johnson et al., 2009;

Thomas & January, 2021). Similarly, the NCII Screening Tools Chart ranks screening assessments on classification accuracy based in part on sensitivity and specificity information: assessments with a sensitivity and specificity above .7 are marked as having "Partially convincing evidence", while those with a sensitivity and specificity above .8 are marked as having "Convincing evidence" (NCII Screening Tools Chart, n.d.).

Achieving Recommended Sensitivity and Specificity

Given the prevalence of such recommendations in practitioner-facing research literature, it is important to know if and when screening assessments can meet these standards. There is some evidence that a high benchmark for sensitivity and specificity, such as .9 and .8 posed by Jenkins & Johnson (2008), is difficult to achieve using administratively and analytically simple means (i.e., a single screening assessment using cut-points to assign risk— an approach commonly employed by schools; Jenkins et al., 2013, Prewett et al., 2012). In a recent study, Edwards et al. (2022) demonstrated that there is a mathematical relationship between predictive validity and classification accuracy, such that, assuming a bivariate normal distribution, all common classification indices can be derived knowing just the correlation between a screening assessment and outcome measure and the cut-points used to denote risk on the screener and non-proficiency on the outcome. Edwards et al. applied this formula to simulated samples with differing base rates, screener cut-scores, and screener-outcome correlations, and found that no screener correlating with the outcome at less than .85 achieved sensitivity of .9 and specificity of .8.

A screener-outcome correlation of .85 is high compared to what is observed in practice, especially in the early grades. Usually, this could not be achieved with a screener measuring a single skill (univariate assessment, e.g., a measure of oral reading fluency alone). In a meta-

analysis of curriculum-based screening assessments administered in kindergarten through 2nd grade, January and Klingbeil (2020) found that predictive correlations between screeners measuring individual components of reading and later reading outcomes ranged from .35 to .83. However, one limitation of Edwards et al. (2022) is that only a set range of screener cut-points were tested (at the 10th, 20th, and 25th percentile of scores.) While appropriate given that researchers tend to use established, publisher-provided cut-scores to determine which students are at risk (Klingbeil et al., 2017), it is known that the placement of the cut-point denoting risk on a screener can have large impacts on classification (Compton et al., 2010; VanDerHeyden et al., 2018). In terms of sensitivity and specificity, raising a cut-point will increase sensitivity at the expense of specificity (identifying more at-risk students while potentially including more False Positives), and decreasing a cut-point will increase specificity at the expense of sensitivity. Thus, it may be possible for a screener correlating with an outcome at less than .85 to achieve recommended levels of sensitivity and specificity if a more optimal cut-point is used than the three selected by Edwards et al.

Value of simulations in educational research

Simulations are computer-generated data sets of random numbers that reflect user-specified parameters. These parameters are typically chosen based on existing large-scale data sets or meta-analyses of target phenomena and then systematically manipulated. Simulated data are useful when researchers wish to study the theoretical effects of phenomena that would be very difficult to observe, manipulate, or isolate in natural data. For example, Edwards et al. (2022) simulated data sets to explore the relationship between predictive validity and classification accuracy in screening because it would be unfeasible to administer many screening assessments with varying correlations with the outcome to samples with a range of base rates (or

to identify these conditions in existing data with any level of precision). Further, simulated data afford a high level of internal validity because all parameters are explicitly specified by the researcher. Thus, simulations are useful for testing mechanistic theories of how a target phenomenon could arise given certain conditions.

Simulations have specifically been used in education research to study how various conditions affect the accuracy and longitudinal stability of alternative approaches to identifying students with learning disabilities. Over 20 years ago, Francis et al. (2005) used simulated data to investigate how measurement error around a cut-point may contribute to unstable identification. More recently, Schatschneider et al. (2016) simulated a series of data sets to compare the one-year stability of alternative approaches to identification across samples with different growth patterns (e.g., fan-spread versus mastery learning growth). As mentioned above, several studies have manipulated correlations among screeners and outcomes to study the impact of predictive validity on classification accuracy (Edwards et al., 2022; van Norman et al., 2019). Simulations are also commonly used to test or validate novel statistical techniques on data that reflect population-level parameters. For example, Wagner et al. (2023) used a simulated data set with parameters based on correlation matrices from three large-scale meta-analyses to test the feasibility of a probabilistic or Bayesian approach to identifying individuals at risk for low achievement in reading. Thus, simulation methodology has been and continues to be used in educational research.

Study 1

The first study reported here uses simulated data to extend the results of Edwards et al. (2022) to provide a more precise picture of what screener-outcome correlation is theoretically required to achieve recommended levels of sensitivity and specificity. Specifically, the first study was guided by the following research questions:

1. Using a single screening assessment, what screener-outcome correlation is required to attain a sensitivity of .9 and a specificity of .8, with any possible cut-point, and does it differ depending on the base rate of non-proficiency in the sample?
2. What is the range of positive and negative predictive power when a screening assessment meets the above recommendations for sensitivity and specificity?

Methods

Simulations. Data were simulated using the mvrnorm() function from the MASS package in R (Venables & Ripley, 2002). The R code and output have been made freely available on the Open Science Framework at the following link: https://osf.io/xwyr8/.

First, sixty-five unique data sets were simulated: one per screener-outcome correlation ranging from .35 to .99 at increments of .01. Each data set comprised scores for one screening assessment and one outcome assessment for a very large sample (n = 10000). This large sample size was chosen in order to reduce noise and reliably identify the range of cut-points producing recommended sensitivity and specificity, as correlations tend to stabilize as the sample size increases (Schönbrodt & Perugini, 2013). The lower bound of tested correlations (r = .35) was selected after evaluating a large first-grade screening data for realistic screener-outcome correlations (see Spencer et al., 2014 for description of data set), and considering the findings of January & Klingbeil (2019)'s meta-analysis.

Second, for each data set, the base rate of non-proficiency was manipulated to be 10, 20, and 25%. This was achieved by defining "non-proficiency" on the outcome as scoring at or below the 10th, 20th, or 25th percentile of scores. The base rates were selected to remain consistent with Edwards et al. (2022), and align with the range of prevalence for dyslexia reported in previous literature.

Third, receiver operating characteristic (ROC) analysis was applied to determine the range of cut-points (if any) that produced recommended levels of sensitivity and specificity for each screener-outcome correlation. ROC analysis can be used to calculate the sensitivity and specificity of a test at every possible decision threshold (Pintea & Moldovan, 2009). The pROC package in R was used to perform the ROC analysis (Robin et al., 2011), using the scores on the screening assessment to predict the binary outcome (proficient or non-proficient).

Analysis

For each simulated data set (i.e., screener-outcome correlation) and base rate, the range of ROC-derived cut-points producing a sensitivity equal to or greater than .9 and a specificity equal to or greater than .8 were identified and reported. Positive and negative predictive power (PPP and NPP) were then calculated using the following formulas:

$$PPP = \frac{(Base\ Rate)(Sensitivity)}{(Base\ Rate)(Sensitivity) + [(1 - Base\ Rate)(1 - Specificity)]}$$

$$NPP = \frac{(1 - Base\ Rate)(Specificity)}{(1 - Base\ Rate)(Specificity) + [(Base\ Rate)(1 - Sensitivity)]}$$

Results

Minimum sufficient correlation. Results are presented in Table 1. The left-most column represents the screener-outcome correlation (i.e., one simulated data set). Each screener-outcome correlation has two rows of information corresponding to the minimum and maximum cut points (in percentiles) that produced recommended levels of sensitivity and specificity. The table is divided into three panels for each of three base rates (10, 20, and 25%), with each panel containing five columns providing the minimum or maximum cut point and the classification indices associated with this screener-outcome correlation/cut point/base rate.

No univariate screener correlating with the outcome at less than .83 achieved recommended levels of sensitivity and specificity (sensitivity >= .9 and specificity >= .8). As seen in Table 1, the screener-outcome correlation needed to achieve these recommended values differed depending on the base rate of non-proficiency in the sample. As the base rate increased from 10% to 20% to 25%, the minimum sufficient screener-outcome correlation increased from .83 to .87 to .88. To demonstrate that the effect of base rate on these population-based indices was not due to changing cut-points across base rate conditions, an additional simulation was conducted using a more traditional screening scenario, in which a screener with a pre-specified cut-point was applied to samples differing in base rates of non-proficiency. A data set was simulated in which the base rate of problems in the sample (i.e., cut-point on the outcome) varied from 10 to 50%, but the same cut-point was always used to assign risk on the screener, with cases below the 20th percentile being flagged as at-risk. The results presented in Table 2 show that sensitivity and specificity again changed with the base rate: as base rate increased, sensitivity decreased and specificity increased.

Predictive power. As seen in Table 1, even when a univariate screening assessment correlated highly with the outcome and achieved recommended levels of sensitivity and specificity, positive predictive power (PPP) was low, particularly when the base rate was low. For example, with a screener-outcome correlation of .9 and a base rate of 10%, any cut-point producing recommended levels of sensitivity and specificity produced a PPP ranging from .35 to .48. This means that 50 to 60% of cases identified as at-risk by the screener were False Positives. At the same level of screener-outcome correlation with a 25% base rate, PPP ranged from .65 to .61, meaning over 30% of positive classifications were False Positives. Negative Predictive Power (NPP) was very high across all conditions, meaning that most negative classifications on the screener were True Negatives. Positive predictive power was also highly impacted by the placement of the screening cut-point. For example, with a screener-outcome correlation of .99 and a base rate of 10% (and a sensitivity and specificity above .9 and .8), positive predictive power ranged from .36 to .91 depending on where the screening cut-point was placed.

Discussion

Study 1 had four main findings. First, it was unlikely for any single screening assessment to reach high levels of sensitivity and specificity, regardless of cut-point or base rate. The minimum sufficient screener-outcome correlation to achieve a sensitivity of .9 and a specificity of .8 was r =.83. As mentioned previously, single indicators of reading ability do not often correlate with reading outcomes this highly (January & Klingbeil, 2020), which may in some cases be the result of low reliability, as reliability sets the upper limit on the possible correlation any screener can have with an outcome (Nunually & Bernstein, 1994).

Second, the sensitivity and specificity of a single screening assessment were impacted by the base rate of non-proficiency in the sample. This finding was unexpected, given that

educational screening literature commonly characterizes sensitivity and specificity as being properties of tests rather than samples that are insensitive to changes in base rate. However, past research in the medical field has demonstrated that the prevalence of a disease can affect sensitivity and specificity when the predictor is not truly binary but exists on a continuum with risk determined using an imposed cut-point, which is most often the case when screening for academic difficulties (Brenner & Gefeller, 1997; Leeflang et al., 2013).

Third, even when a screening assessment reached recommended values of sensitivity and specificity, positive predictive power was low, reflecting the presence of many false positive classifications, particularly when the base rate of non-proficiency in the sample was low. This situation would be untenable for schools with limited resources for providing at-risk students with additional support or intervention.

Finally, the present findings stress the importance of using locally-derived, rather than publisher-issued, cut-points for assigning risk on a screening instrument. As seen in Table 1, the range of cut-points associated with a favorable profile of classification differed depending upon the base rate of problems in the sample.

Study 2

Univariate screening is typically not recommended when screening for risk of academic difficulties. Researchers have suggested that using two or more indicators is preferable (*multivariate assessment*; Catts & Petscher, 2018). In the reading literature, use of a hybrid or "constellation" model (including uniquely predictive strengths, weaknesses, causes, and consequences of reading) to identify reading problems has been shown to result in better predictive validity, classification accuracy, and longitudinal stability compared to a univariate approach (e.g., Compton et al., 2010; Wagner, 2008; Wagner, 2018; Spencer et al., 2014). In fact, some studies employing intensive multivariate screening batteries have been able to achieve near-perfect classification accuracy: in a study of first-grade students, Compton et al. (2006) found that using a classification tree approach and a battery of assessments comprising word identification fluency, phonemic awareness, rapid naming, oral vocabulary, and 5 weeks of progress monitoring data predicted year-end word identification with a sensitivity of 100% and a specificity of 93.5%. While PPP and NPP were not reported, there were few false positives and no false negatives with this approach.

However, in practice, the implementation of screening necessitates balance between efficiency and accuracy. While effective, such an administratively and analytically intensive approach requiring 5 weeks of progress monitoring and use of a classification tree may not be tenable for schools for implement for all students several times per year, and further defies the purpose of universal screening which is intended to be a brief assessment of risk rather than a complete diagnostic work-up (Fletcher et al., 2021). Funding for a screening program and the time and bandwidth of teachers and reading professionals may impact a school's ability to

administer multiple or intensive assessments to every student as a universal screening assessment.

Some researchers have proposed that a gated approach to screening may allow practitioners to achieve recommended levels of sensitivity and specificity while minimizing the administrative burden and cost of screening compared to a "direct route" approach (Compton et al., 2010; vanMeveren et al., 2020; Paly et al., 2022). In gated screening, all students complete a brief initial screening assessment, but rather than immediately qualifying for additional support or intervention, students identified as "at risk" complete follow-up testing. Alternatively, a past year's state test scores may be used as the first gate (van Norman, Klingbeil, & Nelson, 2017). This removes not-at-risk students from the sample while at-risk students complete further assessments to further distinguish True from False Positives, increasing the efficiency of screening as only a subset of students complete multiple assessments. While the idea is pragmatic, the research literature on gated screening in screening for educational difficulties is mixed. In some studies, a gated approach has been shown to reduce False Positives compared to univariate screening (Compton et al., 2010; Klingbeil et al., 2017). However, others have shown that the gated approach increases False Negatives compared to univariate and non-gated multivariate approaches (i.e., decreases sensitivity) compared to univariate and non-gated multivariate approaches (van Norman et al., 2017; Klingbeil et al., 2017; van Norman et al., 2019). However, one study found that the gated approach *increases* sensitivity (vanMeveren et al., 2020). Finally, the reduction in False Positives afforded by gated screening may depend on the strength of the correlation between the measures at each gate, with a lower inter-screener correlation resulting in fewer false positives compared to a higher correlation (van Norman et al., 2019).

Present study

Study 2 aimed to evaluate a two-indicator, or multivariate, approach to screening. Specifically, the second study was guided by the following research question:

1. Using two screening assessments, what is the minimum average screener-outcome correlation between two screeners sufficient to attain a sensitivity of .9 and a specificity of .8, and does it differ depending on the base rate of non-proficiency in the sample or how scores on the two screeners are combined to determine risk?
2. What is the range of positive and negative predictive power when a screening assessment meets the above recommendations for sensitivity and specificity?

Methods

Simulations. The R code and output for Study 2 have also been made freely available on the Open Science Framework at the following link: https://osf.io/xwyr8/.

First, data sets were again simulated using the mvrnorm() function from the MASS package in R (Venables & Ripley, 2002). Each data set comprised scores for one screening assessment and one outcome assessment with a sample size of 300. The screener-outcome correlations for each screener were manipulated to range from .5 to .9, while the correlation between the two screeners was set at .7. These parameters were chosen because there are constraints on what correlation structures are possible to simulate due to the need for a positive definite covariance (correlation) matrix as input. This is logical as it is not possible (without missing data) to have two measures that correlate with each other at .3 but both correlate with the same outcome at .9, for example. With an inter-screener correlation of .7, all matrices with screener-outcome correlations between .5 and .9 are positive definite. For each unique combination of correlations, 100 data sets were generated.

Second, as in Study 1, for each data set, the base rate of non-proficiency was manipulated to be 10, 20, and 25%. This was achieved by defining "non-proficiency" on the outcome as scoring at or below the 10th, 20th, or 25th percentile of scores.

Screening approaches. Four methods of assigning risk (i.e., combining scores from the two screeners) were tested: Both, Either, Sum, and Gated. In the Both condition, cases scoring below the cut-points on *both* measures were flagged as at-risk. Cases scoring below the cut-point on just one measure were considered not at-risk. In the Either condition, cases scoring below the cut-point on *either* screening measure were flagged as at-risk. In the Sum condition, the scores on the two screening measures were simply added to create a sum score. Cases at or below the cut-point were considered at-risk. In the Gated condition, cases scoring below a liberal cut-point on the first screener (see below) were flagged. Only these cases identified as "at-risk" by the first indicator underwent the second "gate" of screening, and only cases scoring below the cut-point on this second screener were considered at-risk.

Cut-points. Unlike Study 1, ROC analysis was not used to identify the range of screening cut-points producing recommended sensitivity and specificity. This is because ROC analysis would require scores from the two screeners to be combined (e.g., with logistic regression), and we wished to compare screening approaches in which the scores on the two screeners remained independent (e.g., the Both versus Either approach). Thus, for the Both, Either, and Sum multivariate conditions, 11 cut-points were tested for each simulation, ranging from matching the base rate to 10 percentile points above the base rate. For example, when the base rate was set at 10%, cut-points on each screener were set the 10th, 11th, …, 19th, and 20th percentile. Cut-points on the two screeners were always equal to one another.

In the Gated condition, the cut-points used to assign risk differed. To compensate for the potential loss to sensitivity observed in previous studies (e.g., van Norman et al., 2019), the cut-point for the screener used as the first gate was very liberal, set at the 80th percentile of scores. In other words, cases scoring in the top 20th percentile were not considered at risk and were not included in the second gate of screening. At the second "gate", the cut point was set to vary between 0 and 10 percentile points above the base rate as in other conditions.

Analysis

For each simulated data set (i.e., screener-outcome correlation), base rate, and approach, sensitivity, specificity, PPP, and NPP were calculated using the following formulas:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Hit\ rate = \frac{TP + TN}{TP + FN + \ FP + FN}$$

$$PPP = \frac{TP}{TP + FP}$$

$$NPP = \frac{TN}{TN + FN}$$

Then, the minimum average screener-outcome correlation between the two screeners and range of cut-points (from the tested range of 11 cut-points) producing recommended classification were identified and reported.

To account correlations existing on a non-interval scale, the correlations between each screener and the outcome were first converted to Fisher's Z units, averaged, then converted back to an average correlation. This was accomplished using the FisherZ() function from the DescTools package in R (Signorell, 2017), which applies the following formula:

$$r\text{-}to\text{-}Z\ conversion:\ z_r = tanh^{-1}(r) = 21log(1-r1+r)$$

Results

Minimum sufficient correlation. Results are presented in Tables 3 and 4. The tables are read the same way as Table 1 except for the left-most column which now represents the combined average screener-outcome correlation between the two screeners calculated using the procedure described in the Analysis section. No combined average screener-outcome correlation less than .77 achieved recommended levels of sensitivity and specificity with the tested range of cut-points. Further, this was only possible with the Either and Sum approaches. While the Both and Gated approaches achieved the recommended level of specificity with any screener-outcome correlation, neither approach ever achieved recommended sensitivity. Additionally, the minimum sufficient average screener-outcome correlation varied across screening approaches. For example, as seen in Table 3, with a base rate of 10%, an average correlation of .77 was sufficient to achieve recommended classification when using the Either approach, while an average correlation of .82 was needed when using the Sum approach.

Within each condition, a limited range of cut-points produced recommended sensitivity and specificity. For example, using the Either approach with an average screener-outcome correlation of .8, recommended classification was achieved in the 10% base rate condition with any cut-point on the two screeners set between the 16th and 18th percentile of scores. With an average screener-outcome correlation of .9, recommended classification could be achieved with any cut-point on the two screeners set between the 10th and 19th percentile of scores. The same was true for the Sum approach (Table 4): within the 10% base rate condition, with an average screener-outcome correlation of .85, any cut-point between 17th and 20th percentile could achieve recommended classification, while with an average screener-outcome correlation of .9, any cut-point between the 12th and 20th percentile could achieve recommended classification.

Predictive power. Tables 3 and 4 further demonstrate that using two screeners which each correlate highly with the criterion outcome and have recommended levels of sensitivity and specificity may still result in an untenable rate of false positives, depending on the screening approach, cut-points, and base rate of problems in the sample. For example, using the Either approach, with a base rate of 10%, sensitivity was .93 and specificity was .933, but PPP was .599—meaning that 40% of cases identified as being at-risk by the screener were False Positives.

Discussion

Study 2 demonstrated that using two indicators in screening decreased the correlation needed from each screener and the outcome necessary to achieve recommended levels of sensitivity and specificity compared to screening with a single indicator, but that this depended on the method of combining scores on the two screeners. If risk was defined as being at risk on *either* of the two screening assessments, it was possible to achieve a sensitivity above .9 and a specificity of .8 with two screeners correlating with the outcome at an average of .77. A screener-outcome correlation of .77 is still on the high end of what has been observed in previous studies in the context of screening for reading problems (January & Klingbeil, 2020). If risk was defined by being below a cut-point on the sum score of the two screening assessments, it was possible to achieve recommended sensitivity and specificity with an average screener-outcome correlation of .82— not much of an improvement compared to using a single indicator, which required a screener-outcome correlation of .83. A gated approach or an approach in which risk was defined based on being at risk on *both* screening assessments could not achieve the recommended sensitivity regardless of screener-outcome correlations, cut-points, or base rates.

General Discussion

Implications for basic and applied research

In the context of screening for academic difficulties, researchers tend to report benchmarks for evaluating the technical adequacy of screening assessments based on sensitivity and specificity. One such recommendation is assessments used to screen for reading problems or dyslexia should have a sensitivity of at least .9 and a specificity of at least .8 (Jenkins & Johnson, 2008). There are two reasons that sensitivity and specificity tend to be emphasized over other classification indices: first, they are ubiquitously reported by test publishers, and second, they are described as being insensitive to base rate, implying that practitioners can assume that the sensitivity and specificity reported by the publisher will be the same when applied in their local context.

The findings of the present study demonstrate that issuing stringent recommendations for only sensitivity and specificity of screening assessments may not be useful, because they are very difficult to achieve using a small battery of indicators with realistic levels of predictive validity. Further, issuing specific guidelines for sensitivity and specificity as indices of screener performance may also be misleading, since the present study demonstrated that sensitivity and specificity are not insensitive to base rate as previously assumed, and that even screeners that are highly correlated with outcomes and have acceptable sensitivity and specificity may still have very low levels of positive predictive power.

Thus, the present findings suggest that sensitivity and specificity should not be presented as the priority metrics in evaluating the technical adequacy of screening instruments, and that indices capturing screening performance across a range of base rates should be in the foreground of such recommendations from researchers to practitioners. At a minimum, the impact of base

rate on sensitivity and specificity should be widely discussed, and it should become standard for screening assessment publishers report not singular values for sensitivity and specificity based on data from a norming sample, but a range of sensitivities, specificities, and predictive power indices observed when applying the instrument to multiple samples differing in base rate.

Implications for schools and school psychological practice

The present findings emphasize both the necessity and difficulty of evaluating the technical adequacy of screening assessments. For school staff or administrators in the position to select a screener, the ability to access and interpret research evidence about the reliability and validity of commercially-available assessments is critical, not only to guide the selection of the best assessment but to have realistic expectations for how a screener will perform when applied in their context. Accessing this information is relatively simple: the Academic Screening Tools Chart maintained by the National Center on Intensive Intervention is a free online resource for comparing the classification accuracy, technical standards (e.g., reliability, predictive validity) and usability (e.g., administration time and cost) of commercially-available screeners in the subject areas of math and reading (NCII, n.d.). For each tool, results are reported for how the screener performed when previously applied to one or more test or norming samples.

Evaluating the results is more difficult. Practitioners consulting the Academic Screening Tools Chart will find a large amount of detailed information which they may not have received any formal training on how to navigate or interpret. The findings of Edwards et al. (2022) and the present study suggest that practitioners should at a minimum attend to the predictive validity and classification accuracy of a potential assessment while noting the level of alignment between the test study/sample and their own context and goals. For example, given the relationship between screener-outcome correlation and classification accuracy demonstrated here, looking at

the predictive validity of a potential assessment may be a good starting point as an index of general screener quality.

However, this estimate alone does not provide much information about how effective the screener is at identifying which students are and are not at risk. While classification accuracy indices are intended to be more useful in this regard, the present study demonstrated that they are not agnostic of contextual features such as the base rate of the target condition in the study sample. Practitioners should thus be aware that the majority of screeners listed on NCII's Tools Chart were tested and normed with samples whose base rate of non-proficiency on the outcome was around 20% (NCII, n.d.). Thus, the reported sensitivity, specificity, and predictive power are not directly generalizable to a new setting with a different prevalence of the target condition that is being screened for.

To get a better sense of how a screener may perform in a new context, it may be useful to use a free online tool created by Edwards and colleagues based on the findings of Edwards et al. (2022) (link: https://qmi-fcrr.shinyapps.io/Correlations_Cut-Points_Classification/). This tool allows users to input a screener-outcome correlation (which can be obtained from the NCII Tools Chart), a cut point on the screener, and the cut point on the outcome (i.e., the base rate if using percentiles), and calculates classification indices given these conditions. Notably, school psychologists may not always be in a position to singlehandedly select a screening assessment for their educational context. However, the above information could aid school psychologists in suggesting or supporting systems-levels improvements in their role as part of the leadership team in an MTSS ecology.

Finally, the finding from the present study that the range of cut-points producing recommended levels of sensitivity and specificity differed depending on the base rate in the

sample brings up past calls from researchers that districts or schools should derive and use locally-normed rather than publisher-provided cut-scores when screening (Rice et al., 2023; Grapin et al., 2017; Schatschneider et al., 2008). However, similar to recommendations for classification accuracy, this suggestion requires more nuance than is typically given. Several recent studies comparing the classification accuracy afforded by locally-derived versus publisher-provided cut scores have found that developing local norms increases screener sensitivity but can substantially decrease specificity depending on the approach used (Rice et al., 2023; Nelson et al., 2017; Grapin et al., 2017).

Thus, a more appropriate recommendation may include the preface that school and districts should consider capacity for intervention before developing local cut scores. School psychologists may be uniquely able to assess and deliver this information depending on the roles they play in their particular school setting. Further, as mentioned in the previous statement, there are multiple statistical approaches to developing local cut-scores based on past student data (e.g., ROC curves, discriminant analysis, logistic regression). While researchers such as Grapin et al. (2017) have stated that each of these approaches can technically be accomplished using common software (e.g., Microsoft Excel), it is unclear whether district and school staff/administrators have access to the training and resources needed to confidently understand the benefits, assumptions, and limitations of these procedures. Further, while school psychologists are often trained in data analysis and the critical consumption of research, they may not have the specific content-area expertise or administrative bandwidth to translate these recommendations into practice. There is a need for further research on the barriers faced and supports needed by school psychologists and other school and district practitioners in this regard; however, one suggestion is that it may be beneficial to seek counsel from outside entities such as screener vendors or

nearby research institution to guide the process of deciding upon and enacting a process for locally-normed cut scores.

Limitations

The present study was limited in several ways. First, the present study relied on methods of determining cut-scores that assume linear data. However, it may be useful to test a classification or regression tree approach which does not carry this assumption. Next, in real-world screening contexts, individuals' abilities change over time, and one possibility is that some False Positive classifications occur when individuals who were truly at-risk at the time of screening are truly not-at-risk at the time of the criterion outcome assessment, due to informal intervention or other changes. A valuable future direction would be to build a growth component into the screening simulations to try to reflect this phenomenon and assess its impact on classification. Finally, the present study was not able to manipulate the intercorrelation between indicators in the two-indicator conditions due to limitations on the simulation methodology. Based on the results of van Norman et al. (2019)'s gated screening simulation, however, in which screeners that were less highly correlated with one another led to a greater reduction in false positive classification compared to screeners with a higher intercorrelation, it would be valuable to test the impact of screener intercorrelation on classification with all of the studied multivariate approaches. A future direction may be to investigate alternative methods of simulating data that would allow a lower screener intercorrelation to be tested.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest. The authors have no relevant financial or non-financial interests to disclose.

Tables

1. Table 1. Cut-points (in percentiles) on screener producing recommended levels of sensitivity and specificity for each screener-outcome correlation, and associated classification indices, across base rates.

Base Rate

		10%					20%					25%				
r	Bound	Cut-Point	Sens	Spec	PPP	NPP	Cut-Point	Sens	Spec	PPP	NPP	Cut-Point	Sens	Spec	PPP	NPP
.83	Max	27	.908	.800	.335	.987										
	Min	26	.900	.810	.345	.986										
.84	Max	27	.918	.800	.338	.989										
	Min	25	.900	.824	.362	.987										
.85	Max	27	.926	.800	.340	.990										
	Min	24	.900	.833	.375	.987										
.86	Max	27	.934	.800	.342	.991										
	Min	23	.900	.845	.393	.987										
.87	Max	28	.943	.800	.344	.992	34	.906	.800	.531	.972					
	Min	22	.900	.858	.413	.987	33	.900	.808	.540	.970					
.88	Max	28	.951	.800	.346	.993	34	.916	.800	.534	.974	38	.908	.800	.602	.963
	Min	21	.900	.870	.434	.987	32	.900	.822	.558	.970	37	.900	.810	.613	.960
.89	Max	28	.960	.800	.348	.994	35	.926	.800	.537	.978	38	.920	.800	.605	.968
	Min	20	.900	.881	.456	.988	31	.900	.836	.579	.971	36	.900	.825	.632	.961

.90	Max	28	.969	.800	.350	.996	35	.937	.800	.539	.981	38	.929	.800	.608	.971
	Min	19	.900	.893	.483	.988	30	.900	.850	.599	.971	35	.900	.839	.650	.962
.91	Max	28	.975	.800	.351	.997	35	.948	.800	.542	.984	39	.940	.800	.610	.976
	Min	18	.900	.904	.510	.988	29	.900	.865	.625	.972	34	.900	.853	.671	.962
.92	Max	28	.981	.800	.353	.997	35	.959	.800	.545	.987	39	.951	.800	.613	.980
	Min	17	.900	.914	.538	.988	28	.900	.879	.649	.972	32	.900	.868	.694	.963
.93	Max	28	.986	.800	.354	.998	35	.968	.800	.547	.990	39	.961	.800	.616	.984
	Min	16	.900	.924	.568	.988	27	.900	.893	.678	.973	31	.900	.882	.718	.964
.94	Max	28	.991	.800	.355	.999	36	.977	.800	.550	.993	39	.970	.800	.618	.988
	Min	15	.900	.936	.608	.988	26	.900	.907	.707	.973	30	.900	.897	.744	.964
.95	Max	28	.996	.800	.356	.999	36	.985	.800	.552	.995	40	.979	.800	.620	.991
	Min	14	.900	.946	.649	.988	24	.900	.921	.741	.974	29	.900	.913	.776	.965
.96	Max	28	.999	.800	.357	1.000	36	.991	.800	.553	.997	40	.987	.800	.622	.995
	Min	13	.900	.957	.699	.989	23	.900	.937	.782	.974	28	.900	.930	.811	.965
.97	Max	28	1.000	.800	.357	1.000	36	.995	.800	.554	.999	40	.994	.800	.623	.997
	Min	12	.900	.968	.760	.989	22	.900	.952	.825	.974	27	.900	.948	.851	.966
.98	Max	28	1.000	.800	.357	1.000	36	.999	.800	.555	1.000	40	.998	.800	.625	.999
	Min	11	.900	.978	.821	.989	21	.900	.969	.879	.975	25	.900	.965	.896	.967
.99	Max	28	1.000	.800	.357	1.000	36	1.000	.800	.556	1.000	40	1.000	.800	.625	1.000
	Min	10	.900	.989	.905	.989	19	.900	.985	.938	.975	24	.900	.984	.950	.967

Note. *r = Screener-outcome correlation; Cut = Cut score in Z-score units, rounded to nearest whole percentile; Sens = Sensitivity; Spec = Specificity; PPP = Positive Predictive Power; NPP = Negative Predictive Power.*

2. Table 2. Effect of base rate on sensitivity and specificity when screener cut-point held constant at 20th percentile of scores.

	Base Rate									
	10%		**20%**		**30%**		**40%**		**50%**	
r	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
.4	.436	.826	.378	.845	.341	.860	.314	.876	.290	.890
.5	.509	.834	.432	.858	.383	.878	.345	.897	.313	.913
.6	.596	.844	.494	.873	.426	.897	.378	.918	.336	.936
.7	.688	.854	.562	.891	.478	.919	.411	.940	.358	.958
.8	.794	.866	.644	.911	.534	.943	.446	.964	.379	.979
.9	.910	.879	.748	.937	.599	.971	.483	.988	.396	.996

Note. *r = Screener-outcome correlation; Sens = Sensitivity; Spec = Specificity*.

3. Table 3. Either approach: Minimum average screener-outcome correlation between two screeners producing recommended levels of sensitivity and specificity for each screener-outcome correlation, and associated classification indices, across base rates.

		Base Rate														
		10%					20%					25%				
Avg *r*	Bound	Cut-Point	Sens	Spec	PPP	NPP	Cut	Sens	Spec	PPP	NPP	Cut	Sens	Spec	PPP	NPP
.77	Max	18	.902	.810	.345	.987										
	Min	18	-	-	-	-										
.78	Max	18	.903	.809	.345	.987										
	Min	18	-	-	-	-										
.79	Max	18	.905	.810	.346	.987										
	Min	17	.901	.823	.362	.987										
.80	Max	19	.928	.800	.341	.990										
	Min	16	.900	.838	.382	.987										
.81	Max	19	.936	.801	.343	.991										
	Min	16	.903	.839	.385	.987										
.82	Max	19	.941	.801	.345	.992	24	.902	.804	.535	.970					
	Min	15	.905	.855	.410	.988	24	-	-	-	-					
.83	Max	19	.949	.801	.347	.993	24	.906	.805	.537	.972	27	.903	.801	.602	.961
	Min	15	.906	.854	.409	.988	23	.902	.819	.556	.971	27	-	-	-	-

.84	Max	19	.956	.801	.349	.994	24	.916	.807	.543	.975	27	.906	.804	.607	.963
	Min	14	.905	.869	.434	.988	22	.901	.834	.577	.971	26	.902	.818	.623	.962
.85	Max	19	.966	.802	.352	.995	25	.940	.801	.542	.981	27	.916	.806	.612	.966
	Min	13	.905	.885	.467	.988	22	.906	.836	.580	.973	25	.901	.835	.646	.962
.86	Max	19	.972	.804	.355	.996	25	.947	.801	.543	.984	28	.940	.801	.612	.976
	Min	13	.910	.885	.468	.989	21	.905	.852	.605	.973	25	.904	.835	.647	.963
.87	Max	19	.980	.804	.357	.997	25	.956	.802	.547	.986	28	.946	.802	.614	.978
	Min	12	.908	.899	.501	.989	20	.906	.867	.631	.974	25	.917	.839	.656	.968
.88	Max	19	.987	.805	.360	.998	25	.965	.804	.552	.989	28	.956	.804	.619	.982
	Min	11	.909	.915	.543	.989	20	.916	.870	.638	.976	25	.929	.843	.664	.973
.89	Max	19	.992	.805	.361	.999	25	.974	.806	.558	.992	28	.965	.807	.625	.986
	Min	10	.902	.930	.590	.988	20	.929	.873	.647	.980	25	.941	.847	.673	.977
.90	Max	19	.997	.805	.362	1.000	25	.982	.808	.561	.994	28	.976	.810	.632	.990
	Min	10	.923	.932	.603	.991	20	.944	.876	.656	.984	25	.954	.851	.682	.982

Note. *r = Screener-outcome correlation; Bound = Minimum and maximum cut-points producing recommended classification accuracy; Cut = Cut-point in percentile units; Sens = Sensitivity; Spec = Specificity; PPP = Positive Predictive Power; NPP = Negative Predictive Power.*

4. Table 4. Sum approach: Minimum average screener-outcome correlation between two screeners producing recommended levels of sensitivity and specificity for each screener-outcome correlation, and associated classification indices, across base rates.

		Base Rate														
		10%					20%					25%				
Avg *r*	Bound	Cut-Point	Sens	Spec	PPP	NPP	Cut	Sens	Spec	PPP	NPP	Cut	Sens	Spec	PPP	NPP
.82	Max	20	.901	.878	.451	.988										
	Min	20	-	-	-	-										
.83	Max	20	.906	.878	.453	.988	30	.901	.850	.601	.972	35	.901	.834	.644	.962
	Min	19	.905	.889	.476	.988	30	-	-	-	-	35	-	-	-	-
.84	Max	20	.915	.879	.458	.989	30	.907	.852	.605	.973	35	.909	.836	.649	.965
	Min	18	.904	.900	.502	.988	29	.902	.863	.622	.972	34	.904	.848	.664	.963
.85	Max	20	.931	.881	.465	.991	30	.917	.854	.612	.976	35	.918	.839	.656	.969
	Min	17	.904	.912	.532	.988	27	.901	.888	.667	.973	32	.902	.874	.704	.964
.86	Max	20	.945	.883	.473	.993	30	.931	.858	.621	.980	35	.931	.844	.665	.973
	Min	16	.903	.923	.564	.988	27	.905	.889	.671	.974	31	.902	.887	.728	.965
.87	Max	20	.958	.884	.479	.995	30	.946	.862	.631	.985	35	.944	.848	.674	.979
	Min	15	.903	.934	.602	.989	25	.902	.913	.722	.974	30	.904	.901	.753	.966
.88	Max	20	.972	.886	.486	.997	30	.959	.865	.639	.988	35	.957	.852	.684	.984
	Min	14	.908	.945	.648	.989	24	.902	.926	.752	.974	29	.905	.915	.780	.967
.89	Max	20	.983	.887	.492	.998	30	.972	.868	.648	.992	35	.971	.857	.694	.989

	Min	13	.909	.957	.699	.990	23	.909	.940	.790	.976	27	.901	.940	.835	.966
.90	Max	20	.992	.888	.496	.999	30	.983	.871	.655	.995	35	.982	.861	.702	.993
	Min	12	.910	.968	.758	.990	21	.901	.963	.858	.975	26	.905	.955	.870	.968

Note. *Avg r = Average screener-outcome correlation between two screeners; Bound = Minimum and maximum cut-points producing recommended classification accuracy; Cut-point = Cut-point on each screener in percentile units; Sens = Sensitivity; Spec = Specificity; PPP = Positive Predictive Power; NPP = Negative Predictive Power.*

References

Brenner, H., & Gefeller, O. L. A. F. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in medicine*, *16*(9), 981-991.

Burns, M. K., VanDerHeyden, A. M., Duesenberg-Marshall, M. D., Romero, M. E., Stevens, M. A., Izumi, J. T., & McCollom, E. M. (2022). Decision Accuracy of Commonly Used Dyslexia Screeners Among Students Who are Potentially At-Risk for Reading Difficulties. *Learning Disability Quarterly*, 07319487221096684.

Catts, H. W., & Hogan, T. P. (2020). *Dyslexia: An ounce of prevention is better than a pound of diagnosis and treatment* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/nvgje

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early Identification of Reading Disabilities Within an RTI Framework. *Journal of Learning Disabilities*, *48*(3), 281–297. https://doi.org/10.1177/0022219413498115

Catts, H. W., & Petscher, Y. (2018). Early identification of dyslexia: Current advancements and future directions. *Perspectives on Language and Literacy*, *44*(3), 33-36.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*(2), 394–409. https://doi.org/10.1037/0022-0663.98.2.394

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., ... & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology*, *102*(2), 327.

Edwards, A.A., van Dijk, W., White, C.M., & Schatschneider, C. (2022). Screening screeners: calculating classification indices using correlations and cut-points. *Annals of Dyslexia, 72*, 445–460 (2022). https://doi.org/10.1007/s11881-022-00261-5

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). https://www.congress.gov/bill/114th-congress/senate-bill/1177

Fletcher, J. M. (2012). Classification and Identification of Learning Disabilities. In *Learning About Learning Disabilities* (pp. 1–25). Elsevier. https://doi.org/10.1016/B978-0-12-388409-1.00001-1

Fletcher, J. M., Foorman, B. R., Boudousquie, A., Barnes, M. A., Schatschneider, C., & Francis, D. J. (2002). Assessment of Reading and Learning Disabilities A Research-Based Intervention-Oriented Approach. *Journal of School Psychology*, *40*(1), 27–63. https://doi.org/10.1016/S0022-4405(01)00093-0

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to the identification of LD: IQ and achievement scores are not sufficient. Journal of Learning disabilities, 38, 98–108. doi:10.1177/00222194050380020101

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-Intervention: A Decade Later. *Journal of Learning Disabilities*, *45*(3), 195–203. https://doi.org/10.1177/0022219412442150

Gaab, N., & Petscher, Y. (2022). Screening for early literacy milestones and reading disabilities: The why, when, whom, how, and where. *Perspectives on Language and Literacy, 48*(1), 11-18.

Gearin, B., Petscher, Y., Stanley, C., Nelson, N. J., & Fien, H. (2021). Document Analysis of State Dyslexia Legislation Suggests Likely Heterogeneous Effects on Student and School Outcomes. *Learning Disability Quarterly*, 073194872199154. https://doi.org/10.1177/0731948721991549

Grapin, S. L., Kranzler, J. H., Waldron, N., Joyce-Beaulieu, D., & Algina, J. (2017). Developing local oral reading fluency cut scores for predicting high-stakes test performance. *Psychology in the Schools*, *54*(9), 932-946.

Individuals with Disabilities Education Act, 20 U.S.C § 1400 et seq (2012).

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., Schiller, E., Blackorby, J., Thayer, S. K., & Tilly, W. D. (2013). Responsiveness to Intervention in Reading: Architecture and Practices. *Learning Disability Quarterly*, *36*(1), 36–46. https://doi.org/10.1177/0731948712464963

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How Can We Improve the Accuracy of Screening Instruments? *Learning Disabilities Research & Practice*, *24*(4), 174–185. https://doi.org/10.1111/j.1540-5826.2009.00291.x

Klingbeil, D. A., Nelson, P. M., Van Norman, E. R., & Birr, C. (2017). Diagnostic Accuracy of Multivariate Universal Screening Procedures for Reading in Upper Elementary Grades. *Remedial and Special Education*, *38*(5), 308–320. https://doi.org/10.1177/0741932517697446

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.

Leeflang, M. M., Rutjes, A. W., Reitsma, J. B., Hooft, L., & Bossuyt, P. M. (2013). Variation of a test's sensitivity and specificity with disease prevalence. *Cmaj*, *185*(11), E537-E544.

Mellard, D. F., McKnight, M., & Woods, K. (2009). Response to intervention screening and progress-monitoring practices in 41 local schools. *Learning Disabilities Research & Practice*, *24*(4), 186-195.

Missouri Department of Elementary and Secondary Education. (n.d.). *Dyslexia screening organizer*. Retrieved October 26, 2022, from https://dese.mo.gov/media/pdf/curr-dyslexia-screening-organizer-by-grade

National Center for Education Statistics. (2019). Nation's report card. *National Assessment of Educational Progress*.

National Center on Improving Literacy. (n.d.). State of Dyslexia. Retrieved October 1, 2022, from https://improvingliteracy.org/state-of-dyslexia

National Center on Intensive Intervention. (n.d.). Academic Screening Tools Chart. Retrieved October 1, 2022, from https://charts.intensiveintervention.org/ascreening

No Child Left Behind Act of 2001, Pub, L, No, 107-110, 20 U,S,C, 6301 et seq, (2002).

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric Theory* (3rd). New York: McGraw-Hill.

Ozernov-Palchik, O., & Gaab, N. (2016). Tackling the 'dyslexia paradox': Reading brain and behavior for early markers of developmental dyslexia: Tackling the 'dyslexia paradox.' *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(2), 156–176. https://doi.org/10.1002/wcs.1383

Petscher, Y., Kim, Y. S., & Foorman, B. R. (2011). * The Importance of Predictive Power in Early Screening Assessments: Implications for Placement in the Response to Intervention Framework. *Assessment for Effective Intervention*, *36*(3), 158-166.

Petscher, Y., & Suhr, M. (2022). Considerations for Choosing and Using Screeners for Students With Disabilities. In C. J. Lemons, S. R. Powell, K. L. Lane, & T. C. Aceves, *Handbook of Special*

Education Research, Volume II (1st ed., pp. 83–96). Routledge. https://doi.org/10.4324/9781003156888-8

Paly, B. J., Klingbeil, D. A., Clemens, N. H., & Osman, D. J. (2022). A cost-effectiveness analysis of four approaches to universal screening for reading risk in upper elementary and middle school. *Journal of School Psychology*, *92*, 246-264.

Pintea, S., & Moldovan, R. (2009). The receiver-operating characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Evidence-Based Psychotherapies*, *9*(1), 49.

Prewett, S., Mellard, D. F., Deshler, D. D., Allen, J., Alexander, R., & Stern, A. (2012). Response to intervention in middle schools: Practices and outcomes. *Learning Disabilities Research & Practice*, *27*(3), 136-147.

Rice, M., Erbeli, F., Truckenmiller, A., & Morris, J. (2023, May 8). Universal Screening in Kindergarten: Validity and Classification Accuracy of Istation's Indicators of Progress–Early Reading. *School Psychology*. Advance online publication. https://dx.doi.org/10.1037/spq0000549

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Sanfilippo, J., Ness, M., Petscher, Y., Rappaport, L., Zuckerman, B., & Gaab, N. (2020). Reintroducing Dyslexia: Early Identification and Implications for Pediatric Practice. *Pediatrics*, *146*(1), e20193046. https://doi.org/10.1542/peds.2019-3046

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). Guilford Press. https://doi.org/10.1080/10888438.2015.1107072

Schatschneider, C., Wagner, R. K., Hart, S. A., & Tighe, E. L. (2016). Using simulations to investigate the longitudinal stability of alternative schemes for classifying and identifying children with reading disabilities. *Scientific Studies of Reading*, *20*(1), 34–48.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Research in Personality*, *47*(5), 609-612.

Signorell et al. (2017). DescTools: Tools for descriptive statistics. R package version 0.99.23.

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. https://doi.org/10.1177/0731948714530967

Streiner, D.L. (2003). Diagnosing Tests: Using and Misusing Diagnostic and Screening Tests, *Journal of Personality Assessment*, *81*(3), 209-219, https://doi.org/10.1207/ S15327752JPA8103_03

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*(4), 522.

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*, *46*(2), 110–120. https://doi.org/10.1177/1534508419857232

Troester, K., Raines, R., & Marencin, N. (2022) Universal Screening Within an RTI Framework Recommendations for Classroom Application. *Perspectives on Language and Literacy,* Winter 2022.

VanDerHeyden, A. M. (2011). Evolving standards of diagnostic accuracy in predicting and avoiding academic failure. In *Assessment and Intervention* (Vol. 24, pp. 59-78). Emerald Group Publishing Limited.

VanDerHeyden, A. M., Burns, M. K., & Bonifay, W. (2018). Is more screening better? The relationship between frequent screening, accurate decisions, and reading proficiency. *School Psychology Review*, *47*(1), 62-82.

VanMeveren, K., Hulac, D., & Wollersheim-Shervey, S. (2020). Universal screening methods and models: Diagnostic accuracy of reading assessments. *Assessment for Effective Intervention, 45*(4), 255-265.

Van Norman, E. R., Nelson, P. M., & Klingbeil, D. A. (2017). Single measure and gated screening approaches for identifying students at-risk for academic problems: Implications for sensitivity and specificity. *School Psychology Quarterly*, *32*(3), 405. https://doi.org/10.1037/spq0000177

Van Norman, E. R., Nelson, P. M., Klingbeil, D. A., Cormier, D. C., & Lekwa, A. J. (2019). Gated Screening Frameworks for Academic Concerns: The Influence of Redundant Information on Diagnostic Accuracy Outcomes. *Contemporary School Psychology*, *23*(2), 152–162. https://doi.org/10.1007/s40688-018-0183-0

Venables, W. N., Ripley, B. D. (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

Wagner, R. K. (2008). Rediscovering dyslexia: New approaches for identification and classification. In G. Reid, A. Fawcett, F. Manis, & L. Seigel (Eds.), *The handbook of dyslexia* (pp. 174–191). Thousand Oaks, CA: Sage.

Wagner, R. K. (2018, December 3). *Why Is It So Difficult to Diagnose Dyslexia and How Can We Do It Better? - International Dyslexia Association*. https://dyslexiaida.org/why-is-it-so-difficult-to-diagnose-dyslexia-and-how-can-we-do-it-better/

Wagner, R. K., Waesche, J. B., Schatschneider, C., Maner, J. K., & Ahmed, Y. (2011). Using response to intervention for identification and classification. In P. McCardle, J. R. Lee, B. Miller, & O. Tzeng (Eds.), *Dyslexia across languages: Orthography and the brain-gene-behavior link* (pp. 202–213). Baltimore: Brookes Publishing.

Wagner, R. K., Moxley, J., Schatschneider, C., & Zirps, F. A. (2023). A Bayesian probabilistic framework for identification of individuals with dyslexia. Scientific Studies of Reading, 27(1), 67-81.