Sensitivity, Specificity, LR+, and LR-: What Are They and How Do You Compute Them?

Ellie W. Edman, DEd, NCSP
Lincoln Intermediate Unit 12
ewedman@iu12.org

Timothy J. Runge, PhD, NCSP
Indiana University of Pennsylvania
trunge@iup.edu

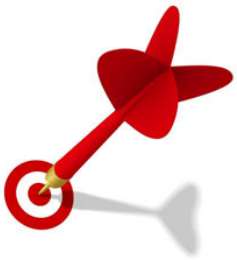


Table of Contents

Necessary Psychometric Qualities of Screening Instruments

A quality screening instrument should have a relatively high success rate of appropriately identifying students categorized in some manner. For example, the STAR Reading (Renaissance Learning, 2014) should do a good job of identifying which students will and will not pass the Pennsylvania System of School Assessment (PSSA) Reading test, the accountability measure utilized in Pennsylvania under the federal No Child Left Behind (NCLB) act.

Since much of the early work on using screening data to predict an outcome comes from medicine, the quality of a screening instrument typically focuses on its ability to predict some negative outcome. In medicine, for example, an electrocardiogram (EKG) is a relatively non-invasive screening of the heart's functionality. The data from the EKG are interpreted to determine the level of risk associated with an irregular heartbeat, damage to the heart tissue, changes in the thickness in the walls of the heart, or electrolyte imbalances in the body. While an EKG alone does not unequivocally rule-in or rule-out these abnormalities, an EKG does a fairly good job of identifying which patients are at an increased risk. If the risk is high enough, the medical team likely will refer for additional, more invasive testing to increase the confidence that a medical condition is present or not.

In the case of academic or behavioral screenings in education, screenings are typically performed to determine which students are in jeopardy of failing the PSSA Reading or experiencing significant social and behavioral challenges. Regardless of the context in which the screening is performed (i.e., education, medicine), judgment of the qualities of a screening instrument requires analysis of a few statistical properties of the screening instrument. These qualities are affected by the selected threshold, or cut score, recorded on the screening instrument.

Sensitivity and Specificity

Sensitivity, the first technical necessity of a screener, is the number of true positives nominated from a particular cut score achieved on a screener (Meehl & Rosen, 1955). For example, a cut score on STAR Reading associated with 80% sensitivity means that 80% of students who fail the PSSA Reading are correctly identified by that particular cut score (e.g., positive screening). Accepted, minimal standards for the sensitivity of a screener are 70% (Glascoe, 2005; VanDerHeyden, 2011).

A second salient psychometric quality of screeners is specificity. Meehl and Rosen (1955) identified specificity as the percentage of true negatives nominated from a particular cut score on a screener. A 90% specificity means that 90% of students who passed the PSSA Reading are accurately categorized by the STAR Reading screening cut score (e.g., negative screening). That is, 90% of students who passed the PSSA Reading also achieved a STAR score above a particular cut score. Glascoe (2005) indicated a specificity level 80% is desirable "so as to minimize overreferrals" (p. 174).

Likelihood Ratios

VanDerHeyden (2011) proposed a simpler solution to the problem of interpreting screening data in an effective and ethical manner. She noted that sensitivity and specificity, while relatively stable despite varying prevalence rates emblematic of educational classification decisions, "offer little information about the value of a [screening] finding for ruling-out or ruling-in a condition" (p. 342). Specifically, sensitivity and specificity do not provide educators with the probability that a student with a particular screening score will fail or pass the PSSA Reading. Predicting the likelihood that a student will pass or fail the high-stakes test is precisely what educators are most interested in calculating from a screener so that appropriate interventions and supports can be delivered for remediation prior to the high-stakes test. Specificity and sensitivity do not provide for such a prediction because the precision of the cut score on the screener is only determined after the high-stakes test is completed. Moreover, cut scores built in sensitivity and specificity only indicate the proportion of all failures and passes on PSSA (sensitivity and specificity, respectively) for that particular score. Sensitivity and specificity do not provide educators with any level of confidence that, for this particular student, he or she will fail or pass the PSSA.

Given the limitations associated with sensitivity and specificity interpretation, educators are called to compute and interpret positive and negative likelihood ratios (LRs). LRs are the likelihood that a certain outcome will occur given a particular performance on a screening measure. LRs, in conjunction with specificity and sensitivity data, provide useful guidance to educators who need to decide whether full evaluation is warranted given a particular screening outcome. As with sensitivity and specificity, LRs change depending on the screening cut score selected.

A positive LR (LR+) is the probability of a positive outcome given a positive screening (i.e., particular score on the screener). This ratio is calculated by dividing the probability of an individual with a positive outcome having a positive screening by the probability of an individual with a negative outcome having a positive screening (VanDerHeyden, 2011). In the context of reading screening, a LR+ relates to the probability that a student who failed the PSSA Reading achieved a particular cut score on a the STAR Reading screener divided by the probability that a student who passed the PSSA Reading achieved the same cut score on the STAR Reading screener. This statistic indicates the odds that the student will fail the PSSA Reading if a particular cut score on STAR Reading is achieved.

Interpretation of LR+s is fairly straightforward with a LR+ = 1 indicating the STAR Reading screener does not predict PSSA Reading failure above random chance. A negative valence to the LR+ indicates a decreased probability, whereas LR+s with a positive valence indicate increased probability of the particular screening performance to predict failing the PSSA Reading. The absolute value of the LR+ represents the magnitude of the probability with LR+s of 1-2 indicating minimal probability; 2-5 indicating small probability; 5-10 indicating moderate probability; and >10 indicating large and conclusive probability (Office of Medical Education Research and Development, Michigan State University, n.d.). Therefore, the larger the magnitude of the positive valence to the LR+, the more accurate the screening cut score is at predicting failing the PSSA Reading.

A negative LR (LR-) is the probability of a negative outcome given a negative screening. This is calculated by dividing the probability of an individual with a positive outcome having a negative screening by the probability of an individual with a negative outcome having a negative screening (VanDerHeyden, 2011). Again, in the context of reading screening, a LR- is the probability of a student who failed the PSSA Reading achieving the cut score on the STAR Reading screening divided by the probability of a student who passed the PSSA Reading achieving the same cut score. This statistic signifies the odds that the student will pass the PSSA Reading if screening performance is at particular cut score.

LR-s range from zero to 1 with values closer to zero representing a stronger likelihood that a STAR screening performance at that particular cut score accurately categorizes the student as failing the PSSA Reading. General interpretative guidelines indicate that an LR- from 0.0 – 0.2 provides relatively high probability that the student will pass the PSSA Reading if the student performs at that particular cut score. An LR- from 0.2 – 0.5 represents a moderate probability that a student will pass the PSSA Reading if the student performs at that cut score. An LR- from 0.5 – 1.0 is interpreted to mean there is a very minimal probability that the student will pass the PSSA Reading if that cut score is achieved. Ideally, an LR- closest to zero is preferable (Office of Medical Education Research and Development, Michigan State University, n.d.); however, a balance between a large LR+ and a small LR- must be negotiated.

Illustration

The following is an illustration of how to calculate and utilize sensitivity, specificity, LR+, and LR- statistics within the context of a commercially-available universal reading screener used to measure annual skill acquisition and predict performance on a high-stakes NCLB accountability test (i.e., PSSA). Additionally, interpretive comments are embedded to help the reader understand how these statistics may be utilized when making data-informed decisions about students and access to tiered academic supports.

Setting and Instrumentation

The Any Area School District (AASD) universally screened all third through sixth grade students in 2012-2013 using the Student Risk Screening Scale (SRSS; Drummond, 1994) and STAR Reading (Renaissance Learning, 2014). The SRSS was completed in October, 2012 and again in January, 2013 per standardized procedures and practices of behavioral screening. Briefly, teachers completed the SRSS by rating each student on seven items using a four-point Likert scale. Previous research provides ample evidence for the reliability and validity of the SRSS as a screener of externalizing problem behaviors (Lane, Kalberg, Bruhn, Mahoney, & Driscoll, 2009; Lane, Little, et al. 2009). The STAR Reading screeners were completed in early September and January per standardized procedures.

STAR Reading data were analyzed from fall and winter screening periods. SRSS scores were entered by classroom teachers into Performance Plus, a secure web-based data warehouse, in November, 2012 and February, 2013. The PSSA Reading assessment was administered in April, 2013.

Archival data from students in third through sixth grades in the AASD were included in the analysis. Publicly available demographic data for each of the four elementary schools were merged with the obtained archived, anonymous data to provide a thumbnail sketch of the district make-up. Demographic data by each of the four elementary schools are summarized in Table 1. Limited demographic data on the sample were available per the approved Institutional Review Board (IRB) protocol. The number of females and males per grade level in the sample are indicated in Table 2. The exact number of students with Individualized Educational Programs (IEPs) for each building was not available to the researchers, although 13.7% of all students in AASD were identified for special education services in that year. Data from all students with disabilities, however, were included in the statistical analyses.

Table 1
Demographic Data for AASD Elementary Schools 2011-2012

School	Grades	Enrollment	% Free / Reduced Meals	% Racial Minority	PSSA Reading Proficiency	PSSA Math Proficiency
A	K-6	490	36.5%	10.1%	75.9%	86.4%
B	K-6	522	42.3%	10.8%	68.7%	78.6%
C	K-6	515	40.6%	8.9%	70.1%	78.5%
D	K-6	513	41.3%	19.9%	66.4%	74.4%

All data, including demographic information and assessment scores, were compiled into a spreadsheet by an AASD technology specialist. The data were de-identified by district personnel before sharing with the researchers per the procedure outlined in the approved IRB protocol.

Table 2
Sample by Sex and Grade

	Total Students	Male	Female
3rd Grade	309	149 (48%)	160 (52%)
4th Grade	294	165 (56%)	129 (44%)
5th Grade	305	157 (52%)	148 (49%)
6th Grade	303	161 (53%)	142 (47%)

Note. Rounding of percentages may result in percentages summing to more than 100%

Utility of STAR and SRSS in Predicting PSSA Reading Performance

To determine the utility of STAR Reading and SRSS fall and winter screenings in predicting PSSA Reading performance, Pearson correlations were computed and logistic regression was performed. Logistic regression analysis explores the predictive power of independent variables on a dichotomous dependent variable. In this case, logistic regression measured the significance of STAR Reading and SRSS in predicting failing or passing the PSSA Reading test.

Local cut scores. To ultimately generate locally-derived cut scores from the universal screening, Receiver Operating Curve (ROC) analyses were completed. Specificity values are graphically represented through a ROC graph. A ROC graph shows true positives on the X axis and false positives on the Y axis for a given predictor and its outcome measure. The diagonal line through the middle of the ROC graph represents random chance of a predictor accurately categorizing the outcome (50% sensitivity & 50% specificity). The area under the curve (AUC) value represents the accuracy of the screening tool and ranges in value from 0.0 (no predictive benefit) to 0.5 (50% accurate predictor) to 1.0 (perfect predictor). AUC values are typically interpreted: 0.0 – 0.4 (worse than random chance), 0.5 – 0.6 (poor), 0.6 – 0.7 (weak), 0.8 – 0.9 (moderate), and 0.9 – 1.0 (excellent).

Separate ROC analyses were completed for the following predictor variables: STAR Reading Fall, grades 3 through 6; STAR Reading Winter, grades 3 through 6; SRSS Fall, grades 3 through 6; and SRSS Winter, grades 3 through 6. Spring screening scores were not included in the analysis because they occurred after the PSSA test was administered. Thus, prediction of early-spring PSSA was not informative using late spring STAR or SRSS.

Sensitivity and specificity values for fall and winter STAR Reading scores at each grade were entered into a spreadsheet. This spreadsheet calculated positive and negative likelihood ratios for scores from each assessment. The researchers used these values, which were derived from district data, to recommend possible local cut scores.

Results

Utility of STAR and SRSS in predicting PSSA reading performance. Significant, strong positive correlations were found between STAR Reading scores and the PSSA Reading test, $r = .758$-$.812$; $p < .001$, at all grades and for each universal screening period analyzed. This means that as a student's STAR Reading test score increases, the predicted PSSA Reading score will also increase. SRSS scores had moderate negative correlations with the PSSA Reading test, with *r* values between -.378 and -.479. As a student's score on the SRSS increases, his or her predicted score on the PSSA Reading test will decrease.

Table 3
Pearson Correlations Between Independent Variables and PSSA Reading Scores

	STAR	SRSS
3rd Fall	.768**	-.369*
3rd Winter	.758**	-.465**
4th Fall	.773**	-.414**
4th Winter	.812**	-.378**
5th Fall	.772**	-.432**
5th Winter	.772**	-.403**
6th Fall	.788**	-.501**
6th Winter	.783**	-.479**

* $p < .05$; ** $p < .01$

Logistic regression was completed to determine the utility of STAR Reading and SRSS in predicting a failing or passing score on the PSSA Reading test. The model contained two independent variables (SRSS and STAR Reading) and the dependent variable of PSSA Reading test. The logistic regression was repeated at fall and winter for each grade level.

Random chance prediction of passing the PSSA test means that we would predict all students to pass the PSSA. Using last year's passing rate of 72% as our only indicator of passing the PSSA Reading the current year, our random chance prediction model (predicting 100% pass rate) would result in 72% of those predictions being accurate. This predictive scenario is the *chance* scenario by which any other prediction models are tested to determine if the predictive model is improved (Berry, 1996).

At each grade level and benchmark period, using a combination of STAR Reading and SRSS increased accuracy in predicting the PSSA Reading from chance. For example, in fall of sixth grade, there was 72% accuracy in predicting a passing score on the PSSA Reading test without considering any predictive measures. When using STAR Reading and SRSS as predictors, there is 80.6% accuracy in predicting a passing score on the PSSA Reading test. General rules of thumb suggest that increasing a prediction model by 5% is considered practically useful above the random chance model (Howell, 2010). As noted in Table 3, all models that included STAR and SRSS improved predictive value over the chance model.

Table 3
Accuracy in Predicting Passing or Failing Score on PSSA Reading with and without Predictors of STAR Reading and SRSS

	Accuracy without STAR and SRSS	Accuracy with STAR and SRSS	Accuracy Change
3rd Grade			
Fall	79.6%	88.0%	+8.4%
Winter	79.6%	86.6%	+7.0%
4th Grade			
Fall	65.9%	82.0%	+16.1%
Winter	66.2%	84.9%	+18.7%
5th Grade			
Fall	66.7%	84.8%	+18.1%
Winter	62.9%	84.2%	+21.3%
6th Grade			
Fall	72.3%	80.6%	+8.3%
Winter	72.6%	82.2%	+9.6%

Note. The accuracy without STAR and SRSS model used the previous year's PSSA Reading passing rate to estimate the accuracy rate of predicting the chance model (i.e., 100% pass rates)

Next, analyses of the strength of each of the predictors (STAR and SRSS) were considered to determine if both or only one of the predictors added predictive value over the chance model. For all but two models (Winter 3rd Grade and Fall 4th Grade), SRSS was not a significant predictor in the model. In other words, in the majority of models, SRSS did not meaningful contribute to predicting PSSA Reading scores above what was already predicted by STAR Reading. STAR Reading scores were significant at $p < .001$ at all grade levels and benchmark periods, providing empirical evidence that STAR Reading is a meaningful predictor of PSSA Reading performance. Additionally, SRSS generally did not meaningfully add to the prediction of PSSA Reading. See Tables 4-7 for results from each grade level analysis.

Table 4
3rd Grade Logistic Regression Variables in the Equation of STAR Reading and SRSS Predicting PSSA Reading

	B	SE	*df*	Sig.	Exp(B)
Fall SRSS	.082	.060	1	.174	1.085
Fall STAR	-.020	.003	1	<.001	.980
Constant	4.615	.881	1	<.001	100.969
Winter SRSS	.180	.072	1	.012	1.197
Winter STAR	-.018	.003	1	<.001	.982
Constant	5.149	1.210	1	<.001	172.332

Table 5
4th Grade Logistic Regression Variables in the Equation of STAR Reading and SRSS Predicting PSSA Reading

	B	SE	*df*	Sig.	Exp(B)
Fall SRSS	.157	.076	1	.038	1.170
Fall STAR	-.017	.002	1	<.001	.983
Constant	6.658	1.169	1	<.001	778.665
Winter SRSS	-.017	.070	1	.809	.983
Winter STAR	-.018	.002	1	<.001	.983
Constant	8.460	1.296	1	<.001	4719.911

Given the results of the logistic regressions, subsequent prediction analyses focused exclusively on STAR Reading predicting PSSA Reading. The lack of robust empirical evidence indicating SRSS meaningfully contributed to the prediction of PSSA Reading was somewhat surprising given contradictory results in other studies (Kalberg, Lane, & Menzies, 2010; Lane, Oakes, Menzies, Oyer, & Jenkins, 2013); however, the data from this sample did not support inclusion of SRSS in the remaining analyses.

Table 6
5th Grade Logistic Regression Variables in the Equation of STAR Reading and SRSS Predicting PSSA Reading

	B	SE	*df*	Sig.	Exp(B)
Fall SRSS	.053	.059	1	.374	1.054
Fall STAR	-.017	.002	1	<.001	.983
Constant	8.614	1.285	1	<.001	5507.573
Winter SRSS	.082	.055	1	.134	1.085
Winter STAR	-.016	.002	1	<.001	.984
Constant	8.652	1.293	1	<.001	5719.859

Table 7
6th Grade Logistic Regression Variables in the Equation of STAR Reading and SRSS Predicting PSSA Reading

	B	SE	*df*	Sig.	Exp(B)
Fall SRSS	.091	.051	1	.076	1.095
Fall STAR	-.011	.002	1	<.001	.989
Constant	5.723	1.016	1	<.001	305.865
Winter SRSS	.004	.052	1	.932	1.004
Winter STAR	-.011	.002	1	<.001	.989
Constant	6.578	1.123	1	<.001	719.329

Locally-derived cut scores. Measures of sensitivity and specificity, along with likelihood ratios were calculated to determine local cut scores and facilitate the probability of passing or failing the PSSA at any given STAR and SRSS score. The first step in this process was ROC curve analysis. AUC calculations were consistently higher for STAR Reading scores compared to SRSS scores from respective periods. This indicates higher predictive accuracy of STAR Reading to PSSA Reading compared to SRSS predicting PSSA Reading.

Table 9
AUC Values for STAR Reading and SRSS Predicting PSSA

	STAR Reading	SRSS
3rd Fall	.925	.758
3rd Winter	.905	.821
4th Fall	.897	.775
4th Winter	.907	.703
5th Fall	.926	.740
5th Winter	.928	.749
6th Fall	.890	.748
6th Winter	.891	.695

Note. STAR Reading AUC values denote how well a higher score on STAR Reading predicts

a passing score (proficient or advanced) on the PSSA Reading test. SRSS AUC values denote how well a higher SRSS score predicts a failing PSSA Reading score.

Sensitivity, specificity, and likelihood ratios were calculated using ROC analysis values. Because it was determined through linear regression that STAR Reading was a strong predictor of PSSA Reading and that SRSS did not meaningfully contribute to the prediction over that which was predicted based on STAR Reading alone, only STAR Reading cut scores are provided. Table 10 lists proposed locally-derived cut scores for AASD at grades 3 through 6 at the fall and winter benchmark periods. Again, cut scores for spring were not calculated due to the spring STAR Reading assessment occurring after the PSSA Reading test.

Table 10
Proposed Local Cut Scores for Predicting PSSA Reading with Fall and Winter STAR Reading

	STAR	Sensitivity (ideal >.70)	Specificity (ideal >.80)	Positive Likelihood Ratio (ideal >2.0)	Negative Likelihood Ratio (ideal <0.2)
3rd Fall	314	.83	.83	4.98	0.20
3rd Winter	366	.75	.87	5.92	0.29
4th Fall	479	.86	.75	3.43	0.19
4th Winter	549	.88	.72	3.13	.016
5th Fall	569	.94	.77	4.00	0.08
5th Winter	622	.91	.77	4.01	0.12
6th Fall	682	.92	.72	3.27	0.11
6th Winter	735	.91	.70	3.04	0.13

Table 11 provides a sampling of multiple STAR Reading scores at winter of 3rd grade level with each score's respective technical qualities. Additionally the table offers a review of the same psychometric qualities of the nationally-derived cut score. Finally, the difference in the number of students identified as being at risk for failing the PSSA Reading at each of the four elementary schools using the various locally-generated cut scores is provided. These data were generated so that school teams could tangibly note the potential change in numbers of students at risk depending on which cut score (national versus local) was used. A positive valence to the difference in the number of students identified by the local versus national cut scores indicates that the locally-derived cut score identified more students as being at risk for failing the PSSA Reading compared to the national cut score. A negative valence to the difference in the number of students identified by the local versus national cut scores indicates that the nationally-derived cut score identified more students as being at-risk for failing the PSSA Reading. Space limitations prevented providing similar comparisons for all other benchmark periods (fall and winter, 3rd through 6th grades).

Table 11
Side-By-Side Comparison of Nationally- and Locally-Derived Winter STAR Cut Scores

STAR SS	Sensitivity	Specificity	LR+	LR-	Difference in Number of Students			
					School A	School B	School C	School D
Nationally-Derived								
352	62%	92%	8.06	0.40	n/a	n/a	n/a	n/a
Locally-Derived								
387	82%	82%	4.45	0.22	10	10	7	9
379	78%	84%	4.72	0.26	7	9	7	9
366	75%	87%	5.92	0.29	4	6	4	4
359	70%	91%	7.63	0.33	1	3	1	1
351	63%	92%	8.06	0.40	0	0	0	0

Note. SS = Standard Score; LR+ = Positive Likelihood Ratio; LR- = Negative Likelihood Ratio

Discussion

Use of STAR Reading in predicting PSSA Reading scores. STAR Reading scores were found to have a stronger correlation to PSSA Reading scores than SRSS, which had a moderate correlation with PSSA. Logistic regression determined that it would be most beneficial to use STAR Reading scores and SRSS scores in isolation when considering intervention planning.

Anecdotal concerns raised by teachers in the AASD suggested that the STAR Reading cut scores were not as precise (accurate) as originally thought. This led to concerns that the STAR Reading test was not identifying students who actually needed intervention. In other words, it was thought that STAR Reading cut scores derived from the national normative sample may have been too low. STAR Reading sets its national benchmark cut-off at the 40th percentile of its normative group for each grade and benchmark period. Further, the 40th percentile rank was based on students' performance on a nationally-normed achievement test, not the PSSA. Given the known differences in rigor among state accountability tests (Kingsbury, Olson, Cronin, Hauser, & Houser, 2004), it is likely more precise to predict STAR performance to the PSSA than STAR performance to some other external criterion. In comparing STAR Reading benchmarks with locally-derived cut scores, STAR benchmark scores were consistently lower.

The locally-derived cut scores summarized in Table 10 were chosen because they had adequate technical properties (e.g., sensitivity, specificity, LR+, and LR-) to predict a passing score on the PSSA Reading test. Table 11 offers an illustration comparing the national and locally-generated (and proposed) cut scores. For example, the STAR Reading national benchmark score for Fall 3rd grade is 352. Using data from AASD, this score corresponds to a sensitivity index of 62% and a specificity index of 92%. These statistics indicate that 63% of 3rd grade students who failed the PSSA Reading test had a Fall STAR Reading score of 352 or lower (sensitivity). Moreover, 92% of 3rd students who passed the PSSA Reading test scored a 352 or higher (specificity). With the publisher's nationally-derived cut score of 352, the district can be confident that 62% of students who fail the PSSA Reading test earned a STAR score of 352 or

lower. However, 38% of students who failed the PSSA Reading scored above the 352 cut score. In other words, using a cut score of 352, the district is under-identifying up to 38% of students who may be in need of reading intervention. This percentage is largely considered unacceptable as it means far too many students are screened to be considered not at-risk for PSSA failure when, in fact, they are at-risk.

In comparison with nationally-derived STAR Reading norms, consider a fall of 3rd grade cut score of 387, which was generated from locally-obtained data. This score has a LR+ of 4.45, considered moderate, but acceptable in education. This means that the district can be moderately confident that a student scoring 387 will fail the PSSA Reading test. A score of 387 has a sensitivity index of 82% and a specificity index of 82%. This means that 82% of 3rd grade students who failed the PSSA Reading test had a score of 387 or lower; however, 18% of 3rd graders who failed the PSSA actually screened higher than a 387. Eighty-five percent of students who passed the Reading PSSA scored a 387 or higher; however, this also means that 15% of students who passed the Reading PSSA scored below 387. Changing the Fall 3rd grade PSSA Reading cut score from 352 to 387 could allow the district to identify an additional 20% of students who are in risk of failing the PSSA Reading test while only falsely identifying 10% more students of potential failure when, in fact, they passed the PSSA Reading (i.e., difference in specificity indices).

Notice that the recommended decision rules above are different from what the STAR program suggests. Two reasons for this disparity: (a) locally-derived norms often tend to be more precise than national norms at predicting performance on an outcome measure (Ferchalk, 2013); and (b) the predicted variable when the STAR was validated was not the PSSA – likely the Group Reading Assessment and Diagnostic Evaluation or Stanford Achievement Test. Therefore, an increasing call for local norming is made in the literature, especially when using these data to allocate students to limited resources.

What does this mean for the district? Setting locally-derived cut scores higher than national cut scores will likely identify more students as potentially being in need of reading intervention; this becomes a matter of resource allocation. Practically speaking, considering locally-derived cut scores should lead to conversations about students. During data analysis meetings, the team should consider both nationally- and locally-derived cut scores. The recommendation would be to look at the nationally-derived cut score and then the locally-derived cut score and discuss the students who fall between those scores. Questions should be raised about these students as to what additional data are available (or should be collected) to more precisely determine risk and assignment to an intervention group.

References

Drummond, T. (1994). *The student risk screening scale (SRSS)*. Grants Pass, OR: Josephine County Mental Health Program.

Berry, D. A. (1996). *Statistics: A Bayesian perspective.* New York, NY: Cengage Learning.

Ferchalk, M. R. (2013). Test nationally, benchmark locally: Using local DIBELS benchmarks to predict performance on the PSSA. Retrieved from ProQuest, UMI Dissertations. (3589383)

Glascoe, F. P. (2005). Screening for developmental and behavioral disorders. *Mental Retardation and Developmental Disabilities Research Reviews, 11*, 173-179. doi: 10.1002/mrdd.20068

Howell, David C. (2010). *Statistical methods for psychology* (7th ed). Belmont, CA; Thomson Wadsworth.

Kalberg, J. R., Lane, K., & Menzies, H. M. (2010). Using systematic screening procedures to identify students who are nonresponsive to primary prevention efforts: Integrating academic and behavioral measures. *Education & Treatment of Children, 33*, 561-584. doi:10.1353/etc.2010.0007

Kingsbury, G. G., Olson, A., Cronin, J., Hauser, C., & Houser, R. (2003). *The state of state standards: Research investigating proficiency levels in fourteen states* (Technical Report). Lake Oswego, OR: Northwest Evaluation Association.

Lane, K. L., Kalberg, J. R., Bruhn, A. L., Mahoney, M. E., & Driscoll, S. A. (2009). Primary prevention programs at the elementary level: Issues of treatment integrity, systematic screening, and reinforcement. *Education and Treatment of Children, 31*, 465-494. doi:10.1353/etc.0.0033

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J. H., Weisenbach, J. L., & Phillips, A. (2009). A comparison of systematic screening tools for emotional and behavioral disorders: How do they compare? *Journal of Emotional and Behavioral Disorders, 17*, 93-105. doi: 10.1177/1063426608326203

Lane, K., Oakes, W. P., Menzies, H. M., Oyer, J., & Jenkins, A. (2013). Working within the context of the three-tiered models of prevention: Using schoolwide data to identify high school students for targeted supports. *Journal of Applied School Psychology, 29*, 203-229. doi: 10.1080/15377903.2013.778773

Office of Medical Education Research and Development, Michigan State University. (n.d.). *Likelihood ratios part 1: Introduction*. Retrieved from: http://omerad.msu.edu/ebm/Diagnosis/Diagnosis6.html

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194-216. doi: 10.1037/h0048070

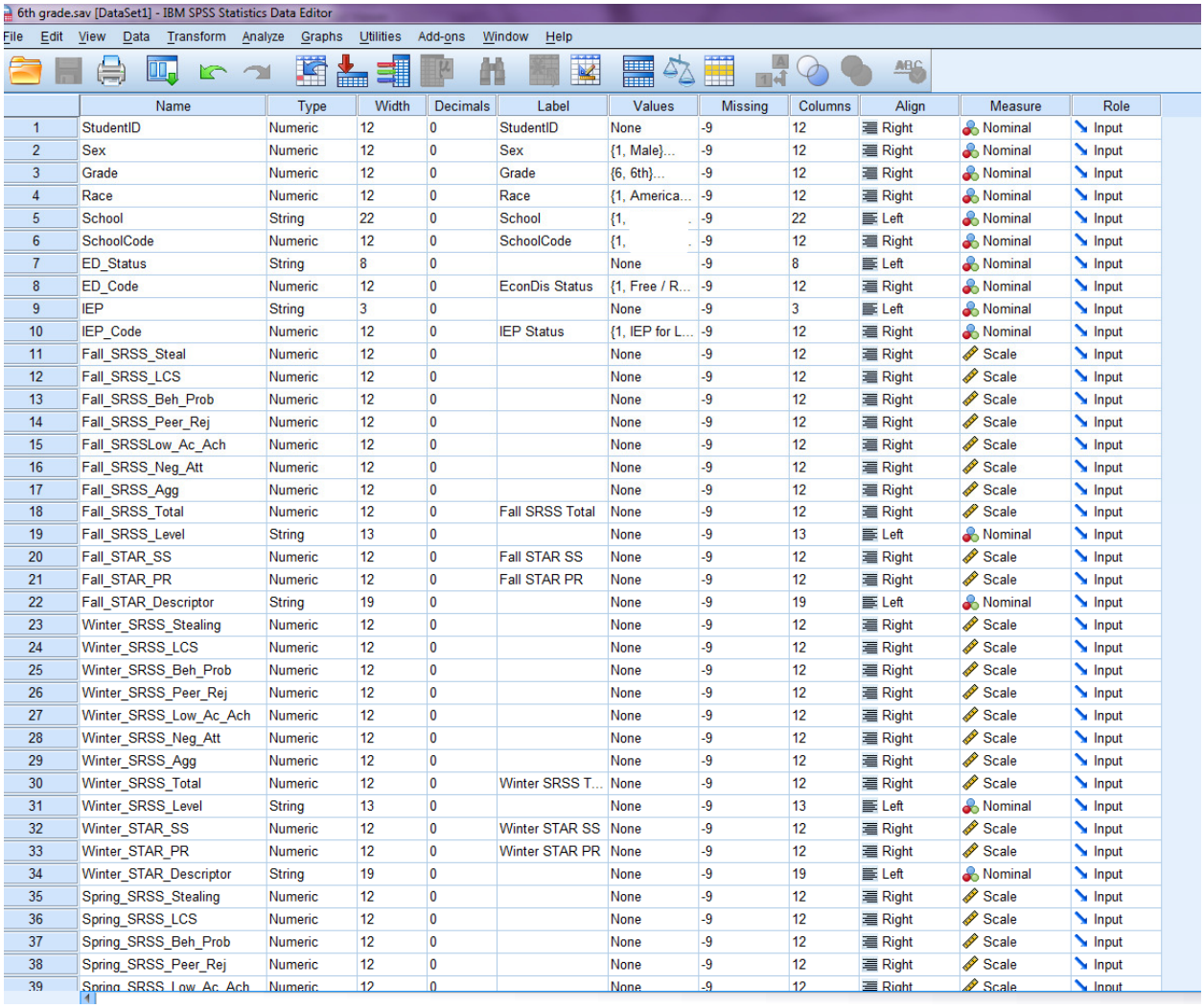
Renaissance Learning. (2014). *STAR Reading Enterprise*. Retrieved from http://www.renlearn.com/sr/default.aspx

VanDerHeyden, A. M. (2011). Technical adequacy of response to intervention decisions. *Exceptional Children, 77*, 335-350.

How to Use the Excel Spreadsheet to Compute LR+ and LR-

Calculating Sensitivity, Selectivity, LR+, and LR-

1. Import data from Excel file into SPSS – likely will need to insert an underscore (_) for all spaces in the variable names listed in excel. SPSS does not like spaces in variables.
2. Make sure all data in SPSS are coded correctly (Variable View).
    a. Use this screen shot as an example to make sure all columns for each data column are correctly coded.


6th grade.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	StudentID	Numeric	12	0	StudentID	None	-9	12	Right	Nominal	Input
2	Sex	Numeric	12	0	Sex	{1, Male}...	-9	12	Right	Nominal	Input
3	Grade	Numeric	12	0	Grade	{6, 6th}...	-9	12	Right	Nominal	Input
4	Race	Numeric	12	0	Race	{1, America...	-9	12	Right	Nominal	Input
5	School	String	22	0	School	{1,	-9	22	Left	Nominal	Input
6	SchoolCode	Numeric	12	0	SchoolCode	{1,	-9	12	Right	Nominal	Input
7	ED_Status	String	8	0		None	-9	8	Left	Nominal	Input
8	ED_Code	Numeric	12	0	EconDis Status	{1, Free / R...	-9	12	Right	Nominal	Input
9	IEP	String	3	0		None	-9	3	Left	Nominal	Input
10	IEP_Code	Numeric	12	0	IEP Status	{1, IEP for L...	-9	12	Right	Nominal	Input
11	Fall_SRSS_Steal	Numeric	12	0		None	-9	12	Right	Scale	Input
12	Fall_SRSS_LCS	Numeric	12	0		None	-9	12	Right	Scale	Input
13	Fall_SRSS_Beh_Prob	Numeric	12	0		None	-9	12	Right	Scale	Input
14	Fall_SRSS_Peer_Rej	Numeric	12	0		None	-9	12	Right	Scale	Input
15	Fall_SRSSLow_Ac_Ach	Numeric	12	0		None	-9	12	Right	Scale	Input
16	Fall_SRSS_Neg_Att	Numeric	12	0		None	-9	12	Right	Scale	Input
17	Fall_SRSS_Agg	Numeric	12	0		None	-9	12	Right	Scale	Input
18	Fall_SRSS_Total	Numeric	12	0	Fall SRSS Total	None	-9	12	Right	Scale	Input
19	Fall_SRSS_Level	String	13	0		None	-9	13	Left	Nominal	Input
20	Fall_STAR_SS	Numeric	12	0	Fall STAR SS	None	-9	12	Right	Scale	Input
21	Fall_STAR_PR	Numeric	12	0	Fall STAR PR	None	-9	12	Right	Scale	Input
22	Fall_STAR_Descriptor	String	19	0		None	-9	19	Left	Nominal	Input
23	Winter_SRSS_Stealing	Numeric	12	0		None	-9	12	Right	Scale	Input
24	Winter_SRSS_LCS	Numeric	12	0		None	-9	12	Right	Scale	Input
25	Winter_SRSS_Beh_Prob	Numeric	12	0		None	-9	12	Right	Scale	Input
26	Winter_SRSS_Peer_Rej	Numeric	12	0		None	-9	12	Right	Scale	Input
27	Winter_SRSS_Low_Ac_Ach	Numeric	12	0		None	-9	12	Right	Scale	Input
28	Winter_SRSS_Neg_Att	Numeric	12	0		None	-9	12	Right	Scale	Input
29	Winter_SRSS_Agg	Numeric	12	0		None	-9	12	Right	Scale	Input
30	Winter_SRSS_Total	Numeric	12	0	Winter SRSS T...	None	-9	12	Right	Scale	Input
31	Winter_SRSS_Level	String	13	0		None	-9	13	Left	Nominal	Input
32	Winter_STAR_SS	Numeric	12	0	Winter STAR SS	None	-9	12	Right	Scale	Input
33	Winter_STAR_PR	Numeric	12	0	Winter STAR PR	None	-9	12	Right	Scale	Input
34	Winter_STAR_Descriptor	String	19	0		None	-9	19	Left	Nominal	Input
35	Spring_SRSS_Stealing	Numeric	12	0		None	-9	12	Right	Scale	Input
36	Spring_SRSS_LCS	Numeric	12	0		None	-9	12	Right	Scale	Input
37	Spring_SRSS_Beh_Prob	Numeric	12	0		None	-9	12	Right	Scale	Input
38	Spring_SRSS_Peer_Rej	Numeric	12	0		None	-9	12	Right	Scale	Input
39	Spring_SRSS_Low_Ac_Ach	Numeric	12	0		None	-9	12	Right	Scale	Input

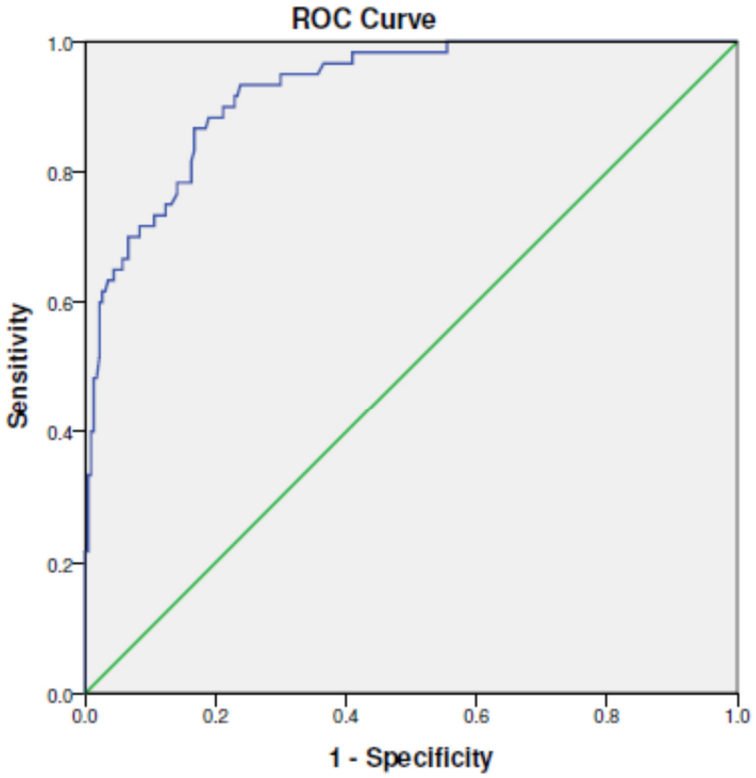
3. Notice that labels were added to some of the variables so that when output is created, the label appears.
4. Values columns need to be entered in by hand using the codebook provided in the original Excel file.

5. A new variable was created at the end called, "PSSA_PASS." This is a dichotomous variable. PSSA performance needs to be dichotomized for many of the analyses computed. This is the easiest way to create this variable and populate it automatically:
    a. Select "Transform" → "Recode into Different Variable"
    b. Highlight "PSSA_Descriptor" variable and move it over to the middle window labeled "Input Variable → Output Variable:"
    c. Label the Output Variable "PSSA_PASS" and type in its label "PSSA Pass / Fail"
    d. Click "Old and New Values"
    e. In next window, enter "Basic" in the "Old Value" value.
    f. Then type in "2" for the "New Value" value
    g. Click "Add"
    h. Repeat e – g for Below Basic (also coded as 2), Proficient (coded as 1), and Advanced (coded as 1).
    i. Click "Continue"
    j. Click "Change" under the right-hand side, Output Variable
    k. Then click "OK"
    l. Double check that a new variable with appropriate coding of 1s and 2s appears in the Data View.
6. Code all variables to include a Missing Value of -9. SPSS considers blank cells as a 0 in some analyses, and this would not be appropriate in most cases.
    a. Go back into the data file to replace all blank cells with -9.
    b. Select "Transform" → "Recode into Same Variables"
        i. Move all variables into the "Variable" window. You may need to do this step separately for Nominal v. Interval data.
        ii. Once all variables are moved to the right side of the window, select "Old and New Values"
        iii. In "Recode into Same Variables: Old and New Values" window, under "Old Value" section of the window, select radial button "System-missing." The within the "New Value" section of the window, enter "-9" for the Value.
        iv. Then select "Add"
        v. Then select "Continue"
        vi. Then select "OK"
        vii. At this point, all the selected variables should now have -9 indicated in cells that were previously empty.
7. Run descriptives and frequencies for all variables to make sure the data are coded correctly. Identify any outliers and determine what to do about them.

ROC Analysis – STAR Data

1. Select "Analyze" → "ROC Curve"
2. Select one "Test Variable" ("STAR_SS" for the appropriate benchmarking period)
3. Select "PSSA_PASS" for "State Variable"
4. Enter "2" for "Value of State Variable." This is stating that we want to see how the "Test Variable" predicts Failing the PSSA (i.e., "2" on "PSSA_PASS").
5. Click Options and a new window will emerge

6. Within that new window, select "Smaller test result indicates more positive test." So what this means is that a smaller STAR score will be used to predict a more positive test (i.e., failing the PSSA Reading)
7. Select "Diagonal reference line," "Standard error and confidence interval," and "Coordinate points of the ROC Curve"
8. Select "OK"
9. The green diagonal in the graph represents random chance that you could predict failing the PSSA (the "State Variable"). So you want the ROC, the blue/purple line to be well above the diagonal. Below is from 3rd grade fall….and the curve is excellent.



10. The Area in the table "Area Under the Curve" is what we interpret. AUCs above 0.7 are considered good; 0.8 or higher is excellent; so below (for 3rd grade fall winter), the AUC of .925 is excellent.

Area Under the Curve

Test Result Variable(s): Fall STAR SS

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.925	.017	.000	.892	.959

The test result variable(s): Fall STAR SS has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

11. Note the subsequent "Coordinates of the Curve" table. These are the data copied back into a separate excel file to compute Specificity, LR+, and LR-. (Excerpt from 3rd Grade Fall STAR is below)

Coordinates of the Curve

Test Result Variable(s): Fall STAR SS

Positive if Less Than or Equal To[a]	Sensitivity	1 - Specificity
70.00	.000	.000
72.50	.017	.000
75.00	.033	.000
80.00	.050	.000
86.00	.067	.000
92.00	.083	.000
97.50	.100	.000
101.00	.117	.000
105.50	.133	.000
110.50	.150	.000
114.00	.167	.000
116.50	.183	.000
119.00	.200	.000
125.50	.217	.000
132.00	.217	.004

12. Open Excel file to populate for computing Specificity, LR+, and LR-. Make sure that the first three columns (Positive If Greater than….; Sensitivity; and 1-Specificity) are all empty.

13. Copy and paste all three columns of the "Coordinates of the Curve" table in SPSS, making sure to get all rows as well. Then paste into the Excel file into the three empty left-most columns.
    a. Notice that once this step is completed, the next columns (D, E, F, and G) will re-calculate automatically.
    b. An excerpt from 3rd Grade Fall STAR is presented below

Positive if Greater Than or Equal to	Sensitivity (% True Positives)	1 - Specificity (False Positive Rate; Type 1 Error)	Specificity (% True Negatives)	1-Sensitivity (False Negative Rate; Type II error)	Positive Likelihood Ratio (probability that the score predicts FAILING PSSA)	Negative Likelihood Ratio (Probability that the score predicts PASSING PSSA)
78.00	0.000	0.000	1.000	1.000		1
80.50	.017	0.000	1.000	0.983	#DIV/0!	0.983333
83.50	.033	0.000	1.000	0.967	#DIV/0!	0.966667
87.00	.050	0.000	1.000	0.950	#DIV/0!	0.95
91.00	.067	0.000	1.000	0.933	#DIV/0!	0.933333
97.00	.083	0.000	1.000	0.917	#DIV/0!	0.916667
103.50	.100	0.000	1.000	0.900	#DIV/0!	0.9
107.50	.117	0.000	1.000	0.883	#DIV/0!	0.883333
127.00	.133	0.000	1.000	0.867	#DIV/0!	0.866667
158.50	.150	0.000	1.000	0.850	#DIV/0!	0.85
175.00	.167	0.000	1.000	0.833	#DIV/0!	0.833333
180.50	.183	0.000	1.000	0.817	#DIV/0!	0.816667
184.50	.200	0.000	1.000	0.800	#DIV/0!	0.8
191.00	.217	0.000	1.000	0.783	#DIV/0!	0.783333
200.00	.233	0.000	1.000	0.767	#DIV/0!	0.766667
212.00	.250	0.000	1.000	0.750	#DIV/0!	0.75
222.00	.267	0.000	1.000	0.733	#DIV/0!	0.733333
224.50	.267	.004	0.996	0.733	61.06667	0.73655
230.50	.283	.004	0.996	0.717	64.88333	0.71981
238.00	.300	.004	0.996	0.700	68.7	0.70307
247.50	.317	.004	0.996	0.683	72.51667	0.68633
258.00	.333	.004	0.996	0.667	76.33333	0.669591
262.00	.350	.004	0.996	0.650	80.15	0.652851
265.50	.367	.004	0.996	0.633	83.96667	0.636111

14. Now all the data needed to create the tables for STAR prediction of **FAILING** PSSA are available.