Evidence for Assessments

Background and Design

Equivalency

Reliability

Decision-Making Accuracy

SpringMath is exclusively provided by Sourcewell Technology, a division of Sourcewell. Sourcewell is a self-funded government organization that partners with education, government, and nonprofits to boost student and community success.

Assessment background and design

- SpringMath measures were built using the science of curriculum-based measurement (CBM).
- Pioneered by Deno & Mirkin (1977), CBM has become the most common assessment used in schools to accomplish screening, to monitor instruction and make mid-stream adjustments, and provide summative evaluations of learning.

Evidence for Assessments

Dr. VanDerHeyden began researching math CBM in 1999.

1999

Developed a set of measures that were the first math CBMs built for kindergarten students and published those data in 2001.

2001

Directed a systematic replication of the 2001 study and included some new measures.

2011

Both studies demonstrated the technical adequacy and utility of these measures which are now used in SpringMath.

Now

STUDENT PRIVACY PLEDGE SIGNATORY

Evidence for Assessments

	Alternate Form *r*	Validity (Brigance)
Count & Circle Number (n = 47)	*r* = .84	*r* = .61
Count & Write Number (n =45)	*r* = .81	*r* = .52
Identify Number & Draw Circles (n = 63)	*r* = .70	*r* = .44

VanDerHeyden, A. M., Witt, J. C., Naquin, G., Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, 30,* 363-382.

Evidence for Assessments

	Alternate Form *r*	Concurrent Validity TEMA	Predictive Validity First Grade CBM Addition	Predictive Validity First Grade CBM Subtraction
Count & Circle Number	r = .84 n = 43	r = .61 n = 44	r = .55 n = 30	r = .55 n = 30
Count & Write Number	r = .71 n = 45	r = .63 n = 45	r = .71 n = 31	r = .51 n = 31
Identify Number & Draw Circles	r = .77 n = 45	r = .58 n = 45	r = .57 n = 31	r = .54 n = 31
Missing Number	r = .87 n = 43	r = .61 n = 43	r = .56 n = 30	r = .52 n = 30
Quantity Comparison with Dots	r = .82 n = 44	r = .41 n = 44	r = .43 n = 31	r = .43 n = 31

VanDerHeyden, A. M., Broussard, C., Snyder, P., George, J., LaFleur, S. M., Williams, C. (2011). Measurement of kindergartners’ understanding of early mathematical concepts. *School Psychology Review, 40,* 296-305.

Evidence for Assessments

- In 2006, VanDerHeyden and Burns conducted the first of three studies that would begin to validate the use of subskill mastery measurement as a reliable, valid, and useful form of assessment for determining response to intervention in mathematics for students in grades 2-5.
- Up to this point, most of the work in mathematics assessment involved trying to create general outcome measures which typically tried to assess multiple skills and model growth over the course of a year.
- VanDerHeyden and Burns believed that more sensitive measurement of skill mastery was necessary to facilitate and inform classwide mathematics intervention.

Burns, M. K., VanDerHeyden, A. M., & Jiban, C. (2006). Assessing the instructional level for mathematics: A comparison of methods. *School Psychology Review, 35,* 401-418.

Evidence for Assessments

- The 2006 study found that fluency scores were more reliable than accuracy scores with reliability values of r = .64 for grades 2 and 3 and r = .88 for grades 4 and 5.
- The standard error of the slope across 4 weeks of progress monitoring was used to calculate the reliability of the slopes for intervention skills with reliabilities of .98, .99, .97, and .98 for grades 2-5.
- Finally, this study demonstrated that fluency scores on foundation tasks could be used to forecast trials to mastery and stronger slope or Rate of Improvement during intervention on subsequent more challenging and complex tasks, which was an empirical validation of the Instructional Hierarchy and powerful evidence that subskill mastery measurement could be used to drive RTI decisions.
- This study also replicated the criteria set forth by Deno & Mirkin (1977) to indicate frustrational, instructional, and mastery level performance in math.
- This was an important contribution because the Deno & Mirkin criteria were interpolated from rates obtained from Precision Teaching implementations (not empirically determined).
- Burns et al. (2006) provided the first empirical validation of Deno & Mirkin’s criteria for math.

Evidence for Assessments

In 2008, 2-week alternate form reliability for measures in grades 2-3 were $r = .71$ and $r = .85$ for grades 4-5.

Decision criteria were tested against the Stanford Achievement Test, 9th edition and found that in grades 2-3 that 34 digits correct per 2 min and 58 digits correct per 2 min in grades 4-5 predicted proficiency on the SAT-9, which basically replicated again the criteria set forth by Deno & Mirkin (1977).

The 2009 study demonstrated, yet again, that early skill proficiency forecasted mastery of more complex, related skills. Empirically derived fluency scores forecasted skill retention, again replicating the functional utility of subskill mastery measures.

VanDerHeyden, A. M., & Burns, M. K. (2008). Examination of the utility of various measures of mathematics proficiency. *Assessment for Effective Intervention, 33,* 215-224. https://doi.org/10.1177/1534508407313482

VanDerHeyden, A. M., & Burns, M. K. (2009). Performance indicators in math: Implications for brief experimental analysis of academic performance. *Journal of Behavioral Education, 18,* 71-91. https://doi.org/10.1007/s10864-009-9081-x

Evidence for Assessments

- In 2010, VanDerHeyden wrote the first of a series of papers articulating a model of academic screening that incorporated local base rates into decision making to improve screening accuracy.
- Specifically, she argued for the use of post-test probabilities to quantify accuracy in local contexts, arguing that base rates of risk would vary across schools and systematically affect assessment accuracies.
- In 2013, she proposed a model, translated from the medical literature of threshold decision making. These concepts are foundational to the value of classwide math intervention as a second screening gate.
- VanDerHeyden demonstrated that once prevalence reaches 50%, even very accurate screenings will not function accurately to rule students out as requiring intervention.
- VanDerHeyden argued for the calculation of post-test probabilities and ruling students out who have a less than 10% chance of failing the year-end test, providing classwide intervention in cases where the probability of failing the year end test ranged from 11-49%, and providing individual intervention to any students with a probability of failing the year-end test greater than 50%.
- This work is the basis for the classwide intervention component of SpringMath.

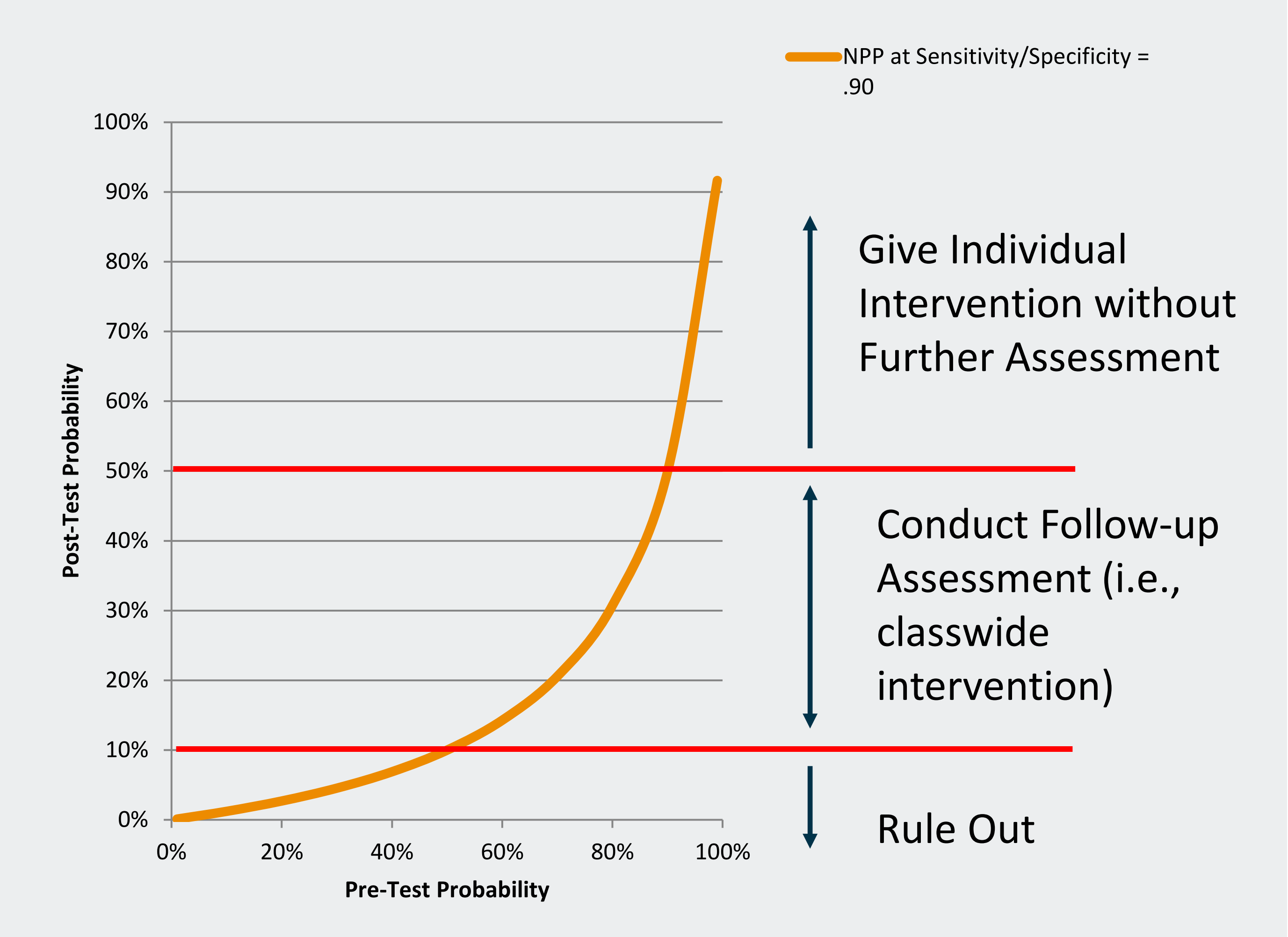
VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. A (2007). Multi-Year Evaluation of the Effects of a Response to Intervention (RTI) Model on Identification of Children for Special Education. *Journal of School Psychology, 45,* 225-256. http://dx.doi.org/10.1016/j.jsp.2006.11.004.

VanDerHeyden, A. M. (2011). Technical adequacy of RtI decisions. *Exceptional Children, 77,* 335-350. https://doi.org/10.1177/001440291107700305

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review, 42,* 402-414.

Evidence for Assessments

- This figure shows the logic of threshold decision making in academic screening.
- At 50% prevalence (pre-test probability), the probability that a student will fail the year-end test even when they have passed the screening is 10% with a very accurate screening measure.
- This is the basis for the classwide intervention in SpringMath.



VanDerHeyden, A. M. (2010). Determining early mathematical risk: Ideas for extending the research. Invited commentary in School Psychology Review, 39, 196-202.

VanDerHeyden, A. M. (2011). Technical adequacy of RtI decisions. Exceptional Children, 77, 335-350.

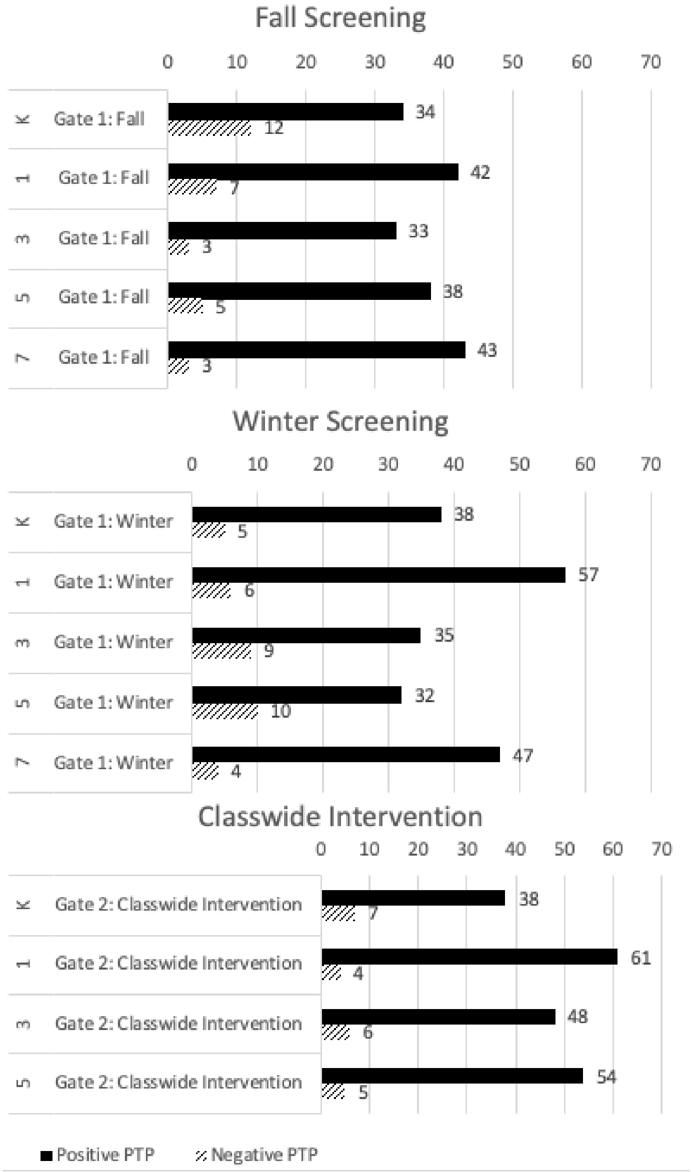
VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. School Psychology Review, 42, 402-414.

Evidence for Assessments

In 2017, VanDerHeyden, Codding, and Ryan applied the threshold model (VanDerHeyden, 2013) to mathematic screening using a variety of measures and found that the subskill mastery measures used in SpringMath outperformed other options and were useful for screening.

In 2019, VanDerHeyden, Broussard, and Burns examined the classification agreement values for fall and winter SpringMath measures and response to classwide math intervention as a second screening gate. They found that classwide math intervention effectively lowered the base rate of risk and permitted superior identification of risk.

Use of classwide math intervention, thus, was empirically validated as an important active ingredient of SpringMath *screening*, both reducing the number of children who required individual intervention but also demonstrating superior negative post-test probabilities.



VanDerHeyden, A. M., Codding, R., Martin, R. (2017). Relative value of common screening measures in mathematics. *School Psychology Review, 46,* 65-87. https://doi.org/10.17105/SPR46-1.65-87

VanDerHeyden, A. M., Broussard, C., & Burns, M. K. (2019). Classification agreement for gated screening in mathematics: Subskill mastery measurement and classwide intervention. *Assessment for Effective Intervention.* https://doi.org/10.1177/1534508419882484

Evidence for Assessments

- SpringMath has a measurement generator that generates equivalent measures for about 135 distinct skills ranging from numeracy to algebra.
- Equivalence is important because any detected change in performance across measurement occasions on the same skill needs to reflect student learning (not changes in difficulty of the measure).
- Potential digits correct per problem type was used to evaluate equivalence and to estimate skill difficulty.
- SpringMath assessments were required to meet rigorous equivalence rules.
- To date, we have generated and tested over 49,000 problems to ensure that our measures function as intended.

Evidence for Assessments

Add 2-Digit Numbers with Regrouping

1 3

$$\begin{array}{r} 88 \\ +\ 34 \\ \hline 122 \end{array}$$

Multiply 2-Digit by 2-Digit with Decimals to the Hundredths

12

$$\begin{array}{r} 0.55 \\ \times\ \ 5.2 \\ \hline 110 \\ 2750 \\ \hline 2.860 \end{array}$$

Use Comparison Method to Solve Systems of Linear Equations

25

$y = 10x + 7$

$y = -4x + 119$

$10x + 7 = -4x + 119$

$14x = 112$

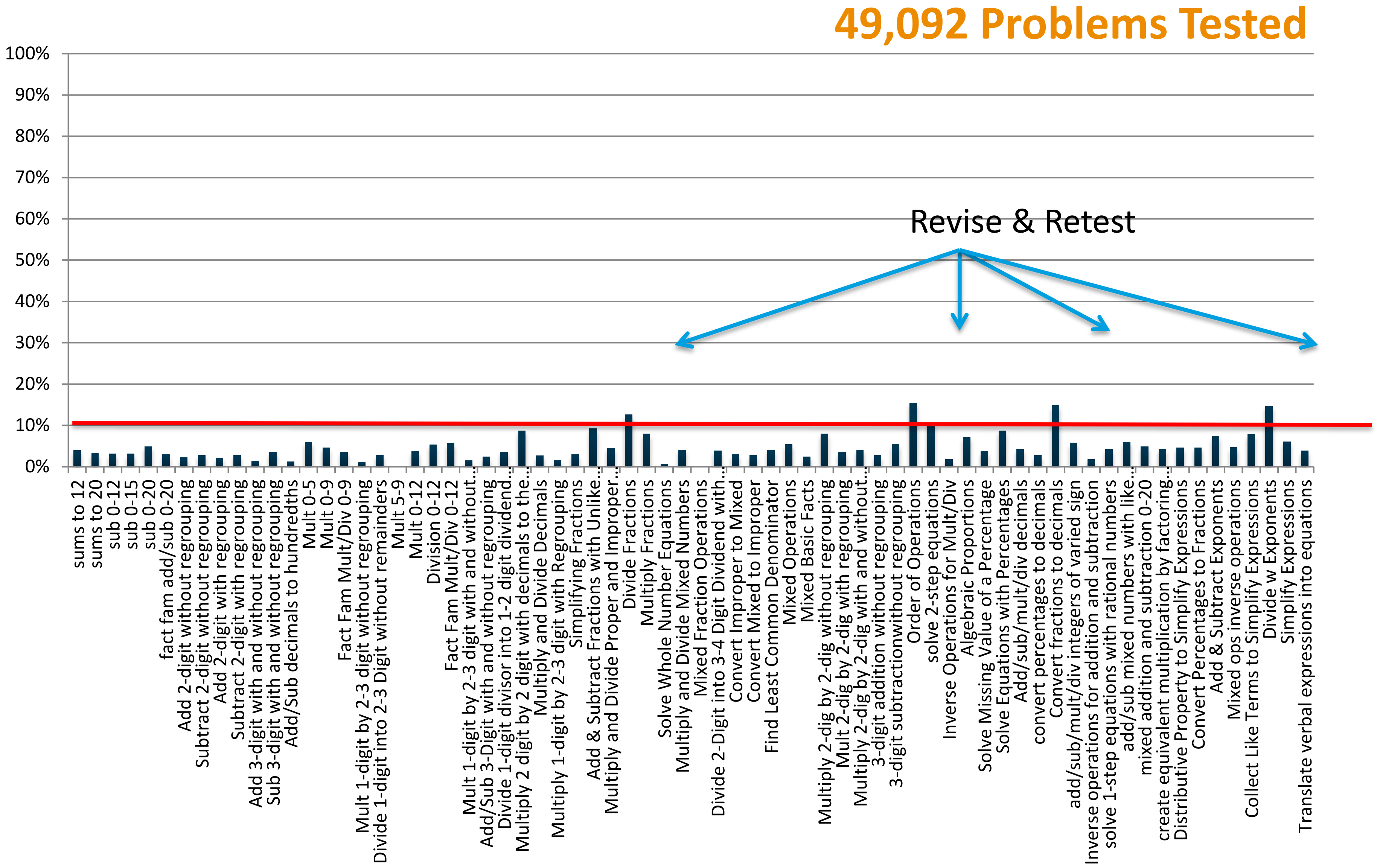
$x = 8$

$y = 80 + 7$

x= y= $y = 87$

More challenging work has more potential digits correct per problem.

Evidence for Assessments



Evidence for Assessments

- SpringMath equivalence testing was a novel approach developed by Dr. VanDerHeyden.
- In 2019, she described this work in an empirical scholarly article in *Assessment for Effective Intervention* for a subset of the measures.
- This study demonstrated, that after testing for the Fall and Winter screening measures (84 measures) the standard deviation of the Mean potential digits correct per problem in a generated problem set was 4% of the Mean digits correct for each specific problem type. In other words, generated problems were equivalent in difficulty according to their potential digits correct.

Evidence for Assessments

Iterations to Equivalence During Phase 1 of the Study			
	Number of Measures that Met Equivalence	Number of Problems Generated and Tested	Percentage of Possible Digits Correct that the SD of Possible Digits Correct Represented*
Round 1	76	23,840	4% (range, 1% to 8%)
Round 2	8	2,850	4% (range, 0 to 9%)

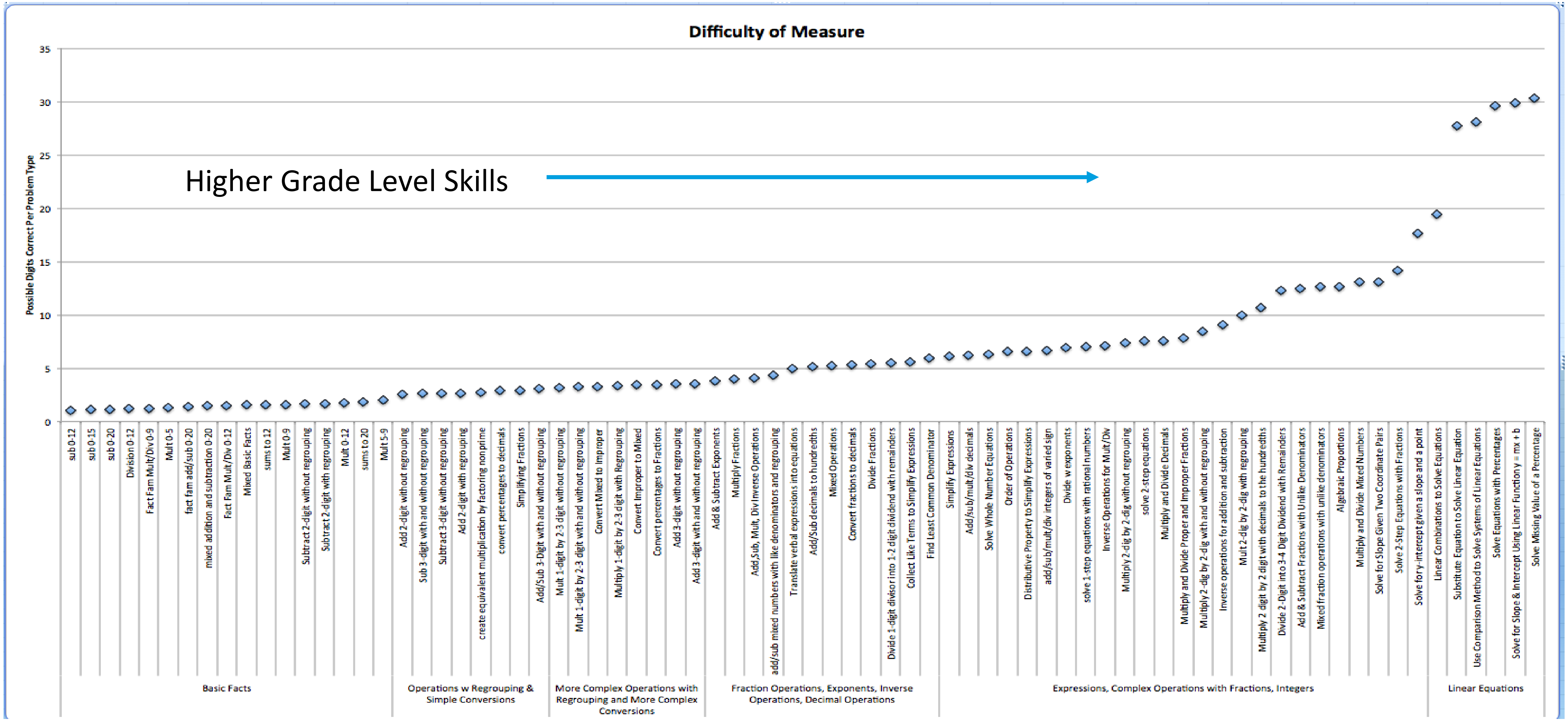
*First, the Mean digits correct per problem for 10 generated problem sets was computed. Next, the SD of the Mean digits correct per problem set for the same problem sets was computed. The SD was divided by the Mean digits correct to indicate what percentage of the Mean digits correct the SD represented. The criterion for stability was that the SD was equivalent to 10% or less of the Mean digits correct.

VanDerHeyden, A. M., & Broussard, C. (2019). Construction and Examination of Math Subskill Mastery Measures. *Assessment for Effective Intervention*. https://doi.org/10.1177/1534508419883947

Evidence for Assessments

- Potential digits correct did function as an indicator of task difficulty as hypothesized by VanDerHeyden.
- The following slide demonstrates the mean possible digits correct across measures, which steadily tracked up reflecting more challenging math tasks across grade levels.

Evidence for Assessments



Evidence for Assessments

- Once equivalence was demonstrated, SpringMath measures were tested in a research study to directly examine reliability.
- With rigorous research controls, 1-week alternate form reliability was tested by administering two generated assessments of the same skill with one week of time between the two assessment occasions. Alternate form reliability ranged from $r = 0.77$ to $r = 0.88$ across grades and assessment occasions. Alternate form reliability was slightly higher at the Winter occasion than the Fall occasion. Mean reliability at Fall was $r = 0.81$ (range, 0.77 - 0.85) and mean reliability at Winter was $r = 0.85$ (range, 0.80 - 0.88).
- It's also important to determine whether the measures could be reliably scored. A total of 1,564 assessments were scored by two independent scorers. Mean IOA across grade levels was 98% (range, 97%-99%). All cases of low agreement (less than 80% agreement) occurred in cases with fewer than 5 attempted answers.
- Drs. Ben Solomon (SUNY at Albany) and Amanda VanDerHeyden collected a large dataset in December of 2019 to quantify the amount of variation in scores that could be attributed to the actual assessment form. These data should be released by winter of 2020.
- What is special about SpringMath measures is that they are not static, but rather are generated as needed. Demonstrating equivalent and reliable scores on generated measures is a novel contribution to the field.

Evidence for Assessments

Grade		*n*	1-Week Alternate Form
Kindergarten	Fall Winter	86 79	*r* = 0.79 (0.69 – 0.86) *r* = 0.80 (0.70 – 0.86)
Grade 1	Fall Winter	79 75	*r* = 0.85 (0.78 – 0.90) *r* = 0.86 (0.78 – 0.91)
Grade 3	Fall Winter	93 91	*r* = 0.82 (0.74 – 0.88) *r* = 0.84 (0.77 – 0.89)
Grade 5	Fall Winter	48 45	*r* = 0.77 (0.62 – 0.86) *r* = 0.87 (0.77 – 0.93)
Grade 7	Fall Winter	41 38	*r* = 0.80 (0.66 – 0.89) *r* = 0.88 (0.78 – 0.94)

VanDerHeyden, A. M., & Broussard, C. (2019). Construction and Examination of Math Subskill Mastery Measures. *Assessment for Effective Intervention*. https://doi.org/10.1177/1534508419883947
Solomon, B., & VanDerHeyden, A. M. (in preparation). G-Study Evaluation of Math Assessment Forms.

The largest-scale series of G studies conducted in math measurement to date

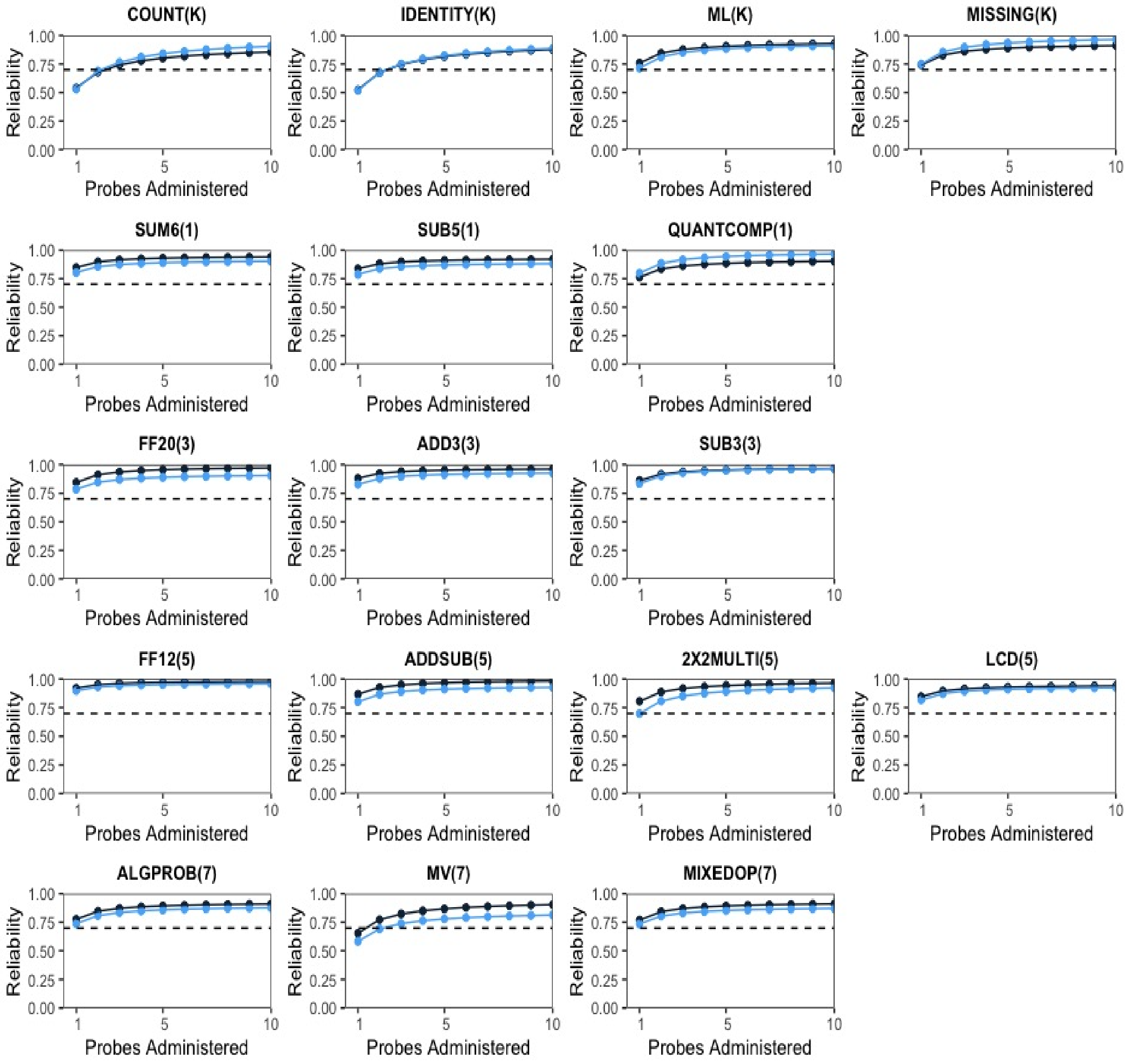
Drs. Ben Solomon (SUNY at Albany) and Amanda VanDerHeyden collected a large dataset in December of 2019 to quantify the amount of variation in scores that could be attributed to the actual assessment form. These data should be released by winter of 2020.

What is unique about SpringMath measures is that they are not static, but rather are generated as needed. Demonstrating equivalent and reliable scores on generated measures is a novel contribution to the field.

This is the largest-scale series of G studies conducted in math measurement to date.

For all measures at all grades, students accounted for the most variance in scores. For 16 of the 17 measures, probe forms accounted for less than 5% of variance. Probe forms accounted for 0% to 4.42% of the variance in scores for the Kindergarten measures, 0.56% to 1.96% for Grade 1, 1.10% to 2.84% for Grade 3, 0.86% to 11.24% for Grade 5, and 0.34% to 2.28% for Grade 7. The measure for which probe forms accounted for 11.24% of variance in scores was Multiply 2-digit by 2-digit Numbers with and without Regrouping in Grade 5.

Thus, the rank ordering of students did not vary based on the probe form. Generalizability coefficients were greater than .7 on the first trial (range, .74 - .92) and .8 (range, .83 - .95) on the second trial for all but three measures. The dependability coefficients followed the same pattern (see Figure). These findings provide rigorous support for the technical equivalence (i.e., stability) of generated measures.



Solomon, B., & VanDerHeyden, A. M. (in preparation). G-Study Evaluation of Math Assessment Forms.