# Symphony Learning Assessment Reliability and Validity

The Symphony Math approach reflects best practices in developmental psychology and cognitive science. Symphony Math stresses substantive and conceptual understanding, rather than rote learning and memorization, so that students can advance quickly to sophisticated problem solving. This approach is reinforced through use of a computer adaptive testing system (CAT) that dynamically selects questions geared to each student's level of knowledge and understanding of the material (for overviews see e.g., Lange, 2007; Wainer, et al., 2000). The targeting of questions to the level of individual students does away with the assumption that "one size fits all;" moreover, efficiency and precision are improved as fewer questions are needed to obtain valid and reliable scores. In addition, by omitting questions that are either too hard or too easy, Symphony Math encourages students to experience testing as a meaningful activity, thereby boosting their motivation to excel. Symphony Math further enhances student motivation by allowing partial credit on many questions, not merely scoring answers as either correct or incorrect (for a discussion, see e.g., Bond and Fox, 2007).

## Assessment Scores

For increased flexibility, student performance is quantified in three ways.

- **Standard scores (SS)** express performance as a number between about 0 and 1100 for students in Kindergarten through Grade 8. The SS define a "vertical scale," meaning that all grades use the same metric and scores are not grade-specific. Thus, a high performing third grader may obtain a higher SS than does a struggling fourth grader. One beneficial property of Symphony Math standard scores is that score differences will have the same meaning across the entire scale. For instance, the standard scores 220 and 250 reflect the same difference in mathematics performance as do SSs of 640 and 670.
- **Grade Equivalents (GE)** re-express standard scores in terms of the grade, plus months of instruction within that grade, as measured against scores typical for students nationwide. GE are written as "year.months;" thus a student' with GE = 7.2 has a SS that is typical for a student in Grade 7 who received two months of instruction in mathematics (typically in November). In this context, the terms "typical score" or "average score" refer to the median value, i.e., the point above and below which fall exactly 50% of all scores. GEs have the advantage of immediately showing students' progress, or lack thereof. Nationwide, the average seventh grader has GE = 7.9 at the end of the school year (i.e., in June) and if a seventh-grader obtained GE = 7.2 this would indicate a lack of progress. Of course, GE = 7.2 reflects advanced performance when obtained by, say, a sixth-grader in January (or any other month for that matter).
- **Percentile Ranks (PR)** express a student's performance relative to that of other students in the same grade nationwide. For instance, a fourth grader whose performance exceeds that of 70% of all fourth graders nationwide is said to score at the 70th percentile. It is equivalent to saying that this student's SS has a percentile rank of 70. Note that since fourth graders score higher on average than do third graders, a fourth grader scoring at the 70th percentile has a higher SS than does a third grader with a score whose PR is 70.

## Test Development

### Item Pool Development

Assessment items were created to measure progress against the Common Core State Standards for Mathematics (see Common Core State Standards Initiative | Home. Web. 13 Feb. 2012. <http://www.corestandards.org/). Three test items were created for each standard from grade kindergarten through eight. The three items for each standard were created at increasing difficulty levels, resulting in an easy, medium and hard level for each CCSE standard. Additional items were created at the pre-kindergarten level to support the assessment of kindergarten students with below grade-level learning. This yielded a total item pool of 900 items. From concept through development of each test question, Symphony Math reflects considerable in-house expertise, as well as that of masters and doctoral level consultants in curriculum and technology design.

### Item Types

The Common Core State Standards present some unique objectives, which in turn necessitated the development of innovative item types. Specifically, the CCSS emphasize a deeper level of understanding and add new skills such as fluency. In order to measure these stated learning goals, we created a variety of item types that challenge students beyond the more superficial responses required by multiple-choice only assessments. The table below details the different item types used by the Symphony Screener and Benchmarker.

| Item Type | Definition |
|---|---|
| Multiple Choice | Select answer from two or four choices. |
| Fill in the Box | Select answer that completes empty box in equaation or sentence. |
| Match | Match 3 or 4 answer choices with corresponding targets. |
| Multiple Correct | Select 2-5 correct answers out of a total of 6 anser choices. |
| Short Answer | Use numeric keypad to enter numeric response. |
| Fluency | Answer under time constraint. |
| Create a Line or Point on a Grid | Place a point on a number line or coordinate grid.<br>Draw a line on a coordinate gride |

In addition to the item types detailed above, students are also given virtual tools to use in solving problems. For example, in some items students must use an onscreen ruler or protractor to measure lengths and angles.

### Pilot Testing

The questions on Symphony Math all survived rigorous pilot testing in agreement with the requirement outlined in the standards for educational and psychological testing provided by the American Educational Research Association. A series of pilot studies were performed using data of over 29,000 students from 9 states, including California, Connecticut, Florida, Illinois, Massachusetts, New York, Rhode Island, Tennessee, and Texas. Throughout these studies, Rasch measurement (see e.g., Bond and Fox, 2007; Custer, Omar, and Pomplun, 2006; Jungnam, et al., 2009) was used to evaluate, calibrate and select the test questions. In particular, ambiguous questions or questions that failed to

consistently reflect knowledge or insight into mathematics were identified and omitted. Also, items biased with respect to demographic subgroups of students were rejected.

As is recommended in the literature (e.g., Baumer et al., 2009), all piloting was computer based using a traditional process?, i.e., form use a fixed set of items, as well as CAT-based tests, to create a vertical scale covering Kindergarten through grade 8. This scale was created by presenting lower-grade questions to students in higher grades. Items targeted to higher grades were also administered to students in lower grades. In this case, however, care was taken not exceed the item – grade difference by more than one level. The psychometric properties of the different question types mentioned earlier were evaluated in detail (Lange, Stevens, and Schwartz, 2011).

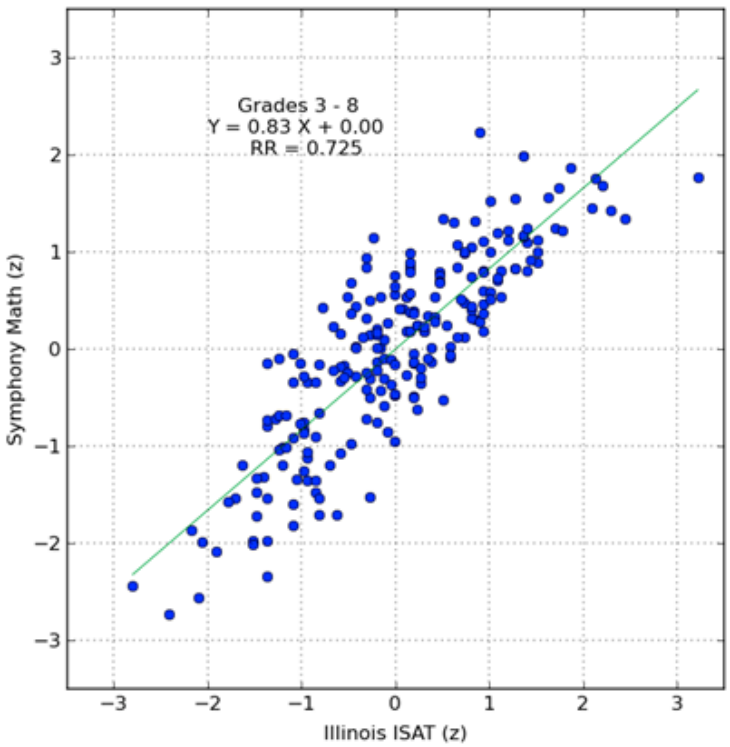
### Reliability

This section focuses on the Benchmark form of the Symphony Math test, which consists of no more than 25 questions. Among the approximately 29,000 students in the calibration sample, a subset of about 15,000 students across grades K through 8 completed 20 to 25 items. The reliability within each of these grades exceeds 0.90, whereas the score reliability across all grades is 0.97 .

### Validity

The validity of the Screener and Benchmarker derives from three independent sources. First, content validity was ensured by a panel of 15 teachers who reviewed each item to confirm whether it was aligned to the designated CCSS or not. Rejected items were revised to address the concerns of the panel.

**Figure C.a:** Correlation Between Symphony Math and ISAT Across Grades 3 Through 8 Based on 2011 Illinois Student Data (N = 218)



Second, the test validity of Symphony Math is addressed through the use of Rasch scaling (Bond and Fox, 2007; Lange, 2007). Valid measurement requires that all items should form a single difficulty hierarchy; this was enforced through item selection and rewriting. Moreover, the item hierarchy was constructed to be invariant across subgroups, thereby ensuring a uniform score interpretation and an absence of bias. To be included in the final pool, items had to demonstrate an absence of bias, i.e., questions were rejected when equally proficient members of different student groups (boys vs. girls, or white vs. black vs. Hispanic students, etc.) had unequal chances of answering questions correctly.

Thirdly, it was possible to study convergent validity by comparing some students' Symphony Math scores to scores obtained on another widely accepted mathematics test. Illinois students in grades 3 through 8 are required to take the vertically scaled ISAT test (see, http://www.isbe.state.il.us/assessment/isat.htm), which includes a shortened form of Pearson's nationally normed SAT-10. Mathematics sub scores on the ISAT were available for 218 students who also completed a pilot form of Symphony Math. As is illustrated in Figure C.a, the correlation between Symphony Math (Y-axis) and the ISAT (X-axis) scores was 0.83, which explains about 73% of the variance. (Note: In the figure the test results are expressed as z-scores). The high level of agreement with students' scores on an established test of mathematical achievement strongly supports the validity of Symphony Math.

### National Norms

National norms were computed for Symphony Math in Grades 4 and 8 based on a two-step linking process.

1. **Symphony Math to New York Mathematics Assessment**. In 2011, a sample of 279 students completed the New York Mathematics Assessment as well as the Symphony Math test. Again supporting convergent validity, the correlation between these two sets of tests scores was 0.78. Using equipercentile equating, Symphony standard scores could thus be expressed on the same scale as the New York Mathematics Assessment.

2. **New York Mathematics to NAEP.** New York's mathematics performance in grades 4 and 8 on the latest NAEP test (given in 2009) can be found via NAEP's Data Explorer at http://nces.ed.gov/nationsreportcard/naepdata/. This site provides means as well as the 10th, 25th, 50th, 75th, and 90th score percentiles nationwide, as well as for each of the 50 states. Assuming normality, New York's test results could thus be transformed into approximate NAEP scores.

**Figure C.b:** Symphony Math Standard scores in Grades 4 and 8 as a Function of Their Equated NAEP scores.

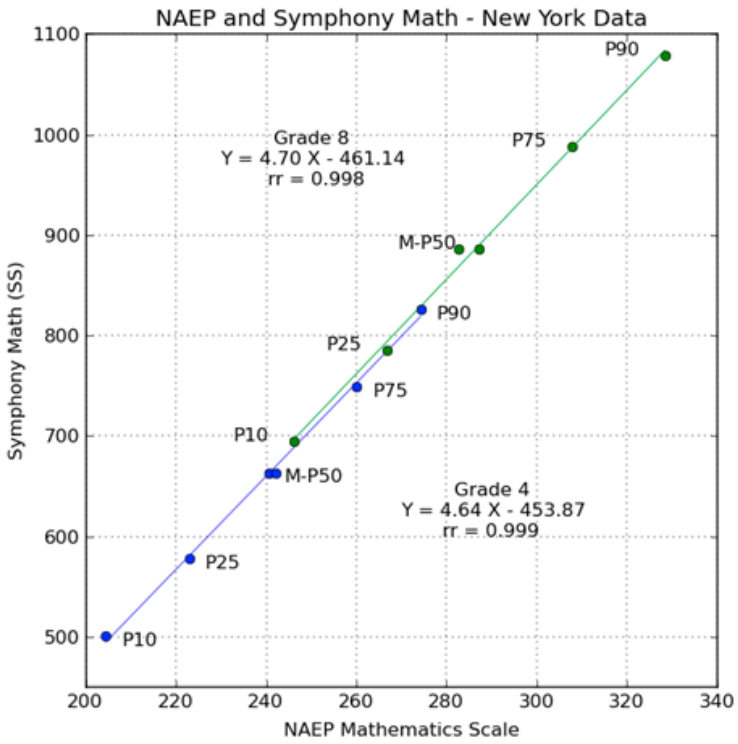



Figure C.b shows the relation between six points in the estimated NAEP and Symphony Math standard scores: The mean (M), P10, P25, P50, P75, and P90. It can be seen that the least-squares regression lines values fit very well, explaining over 99% of the variation among these statistics in the two tests. Most importantly, the high end of the regression line for grade 4 nearly coincides with the lower part of that for grade 8. In other words, the same basic linear relation is carried forward from grade 4 to grade 8. Hence, as is the case for temperatures measured in Centigrade vs. degrees Fahrenheit, Symphony Math and NAEP scores quantify the same concept, as either one is a linear transformation of the other. Using this transformation, national percentiles can be computed for each Symphony Math score. Note that Figure C.b also illustrates the overlap in performance of 4th and 8th graders.

In every aspect, and by every measure, Symphony Math proves an effective system of math assessment and intervention for students from pre-k through 8th grade. Data supports the validity of Symphony Math, showing it strongly correlated to NAEP and state tests like the ISAT. Analysis of scoring data also proves Symphony Math a reliable, consistent assessment across demographic groups. Thus, Symphony Math program provides school districts, schools, classroom teachers and resource professionals with the means to not only efficiently assess and identify students at risk for math failure, but to track and support efforts to bring students' math proficiency in line their peer group.

## References


American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Baumer, M., Roded, K., & Gafni, N. (2009). Assessing the equivalence of Internet-based vs. paper-and-pencil psychometric tests. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.

Bond, T. G., & Fox, Ch. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.). London: LEA.

Custer, M., Omar, M.H., and Pomplun, M. (2006). Vertical Scaling with the Rasch Model Using Default and Tight convergence Settings with Winstepts and Bilog-MG. Applied Measurement in Education, 19, 133-149.

Jungnam, K., Lee, W-C, Kim, D-I, Kelly, K. (2009). Investigation of Vertical Scaling Using the Rasch Model. Paper presented at the annual meeting of the National Council on Measurement in Education. San Diego, CA.

Gorin, J. S., Dodd, B. G., J. Fitzpatrick, S. J., and Shieh , Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. Applied Psychological Measurement, 29, 433-455.

Lange, R. (2007). Binary Items and Beyond: A Simulation of Computer Adaptive Testing Using the Rasch Partial Credit Model. In: Smith, E. and Smith, R. (Eds.) Rasch Measurement: Advanced and Specialized Applications. Pp. 148-180, Maple Grove, MN: JAM Press.

Lange, R., Stevens, D., and Schwarz, P. (2011). Some Surprising Dynamics of "Technology Enhanced" Item Types: Taking Additional Time is Associated with Increased Student Performance on Some Types of Items, while Decreasing Performance on Others. International Association for Computerized Adaptive Testing Conference. Pacific Grove, California, October 3-5.

Loyd B. H., & Hoover, H.D. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 17, 179-193.

Pomplun, M., Omar, H., & Custer, M. (2004). Comparison of WINSTEPS and BILOG-MG for vertical scaling with the Rasch model. Educational and Psychological Measurement, 64, 600-616.

Wainer, H., Dorans, N.J., Flaugher, R., Mislevy, R.J., Green, B.F., Steinberg, L., and Thissen, D. (2000). Computerized Adaptive Testing: A Primer (Second Edition). Hillsdale, NJ: Lawrence Erlbaum Associates.