Technical Report # 1402

Criterion Validity Evidence for the easyCBM©

CCSS Math Measures:

Grades 6-8

Daniel Anderson

Brock Rowley

Julie Alonzo

Gerald Tindal

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Note: Funds for this data set used to generate this report come from a federal grant awarded to the UO from the Institute of Education Sciences, U.S. Department of Education: Developing Middle School Mathematics Progress Monitoring Measures (R324A100026 funded from June 2010 - June 2014).

Abstract

The easyCBM© CCSS Math tests were developed to help inform teachers' instructional decisions by providing relevant information on students' mathematical skills, relative to the Common Core State Standards (CCSS). This technical report describes a study to explore the validity of the easyCBM© CCSS Math tests by evaluating the relation between students' scores on these measures in Grades 6-8, and the Stanford Achievement Test, 10th edition (SAT-10). High correlations between the two would provide evidence for the validity of the easyCBM© CCSS Math Measures, while a low relation would provide evidence that the measures are targeting different constructs and/or functioning differently. We explore the relation between the measures using both correlational and regression analyses. Results suggest a high relation between the measures, adding to the validity evidence for the easyCBM© CCSS Math measures.

Criterion Validity Evidence for the easyCBM© CCSS Math Measures: Grades 6-8

Response to intervention (RTI) has multiple purposes, including monitoring the progress of all students, identifying those who may be in need of additional supports, evaluating the effect of interventions over time, and potentially identifying students for placement in special programs (Fuchs & Fuchs, 2006). Educators working within RTI administer benchmark measures to screen students for their risk for future low achievement and progress monitoring measures to document and evaluate the achievement trends of students performing below expectations. The validity of RTI hinges upon the technical adequacy of the measures used, the subsequent interpretation of the data by the educators, and the inferences drawn informing the final decision made. As Messick (1995) notes, it is the decisions that ultimately are or are not valid. Yet, the validity of decisions relies heavily upon the aforementioned components.

The likelihood of valid instructional decisions being made within RTI depends, in part, upon the extent to which students' scores accurately reflect their knowledge and skills on the target construct. The purpose of this technical report is to examine this relation for the easyCBM© CCSS Math tests in Grades 6-8. All easyCBM© CCSS Math measures are specifically designed for use within RTI, with three alternate test forms of equivalent difficulty developed in each grade for seasonal benchmark screening (fall, winter, spring), and a series of 10 alternate forms available for follow-up progress monitoring. For a complete description of the development of the tests, see Anderson, Irvin, Patarapichayatham, Alonzo, and Tindal (2012) and Anderson, Alonzo, and Tindal (2013a).

Validity is not an all-or-nothing property, and evidence of the validity of a test for an intended use cannot be documented in a single study. As evidence over multiple studies accumulates, inferences of scores being representative of the underlying trait are strengthened.

This study represents the first empirical investigation into the validity of the easyCBM© CCSS Math tests for use within RTI. We collect concurrent criterion validity evidence of the measures by examining the relation between scores on these measures and scores on the Stanford Achievement Test, 10th Edition (SAT-10). Both tests were designed to measure a similar construct, but the SAT-10 has documented validity evidence (Pearson, 2004). If a high relation is found, our confidence that easyCBM© accurately represents the underlying trait is increased. Accurately representing the intended construct does not guarantee valid decisions will be made, but it can perhaps be viewed as a prerequisite. If a low relation is found, it may suggest that the easyCBM© CCSS Math measures target different mathematical concepts, and/or are functioning differently than the SAT-10. Follow-up investigations may then reveal whether revisions are necessary.

Methods

Participants and Procedures

A random sample of students within one school in the Pacific Northwest participated in this study. The sample size necessary to achieve power of at least 0.8 was calculated *a priori*, holding alpha at .05 with an effect size (Cohen's f^2) of 0.4. With a single predictor, a sample size of 22 would be sufficient to detect a significant effect 80% of the time. We purposefully oversampled to ensure adequate power, randomly selecting 65 students from each grade. The random selection was not stratified by any student subgroups. We used the random number generator function in Excel to assign random values to each student in each grade, and then sorted the file by the randomly-generated number and selected the first 65 students. Table 1 reports the sample demographics for students included in the study. Note that prior to analysis, special education (SPED), English language learner (ELL), and free or reduced price lunch

eligibility were collapsed into dichotomous yes/no categories. Ethnicity was collapsed into a White/non-White variable. The more nuanced categories are presented for descriptive purposes only. Further, because of the small representation within each target group, demographic variables were included only on an exploratory basis.

All students in the district were administered the winter easyCBM© CCSS Math benchmark. Within one week of taking the benchmark, the randomly selected students described above were administered the SAT-10. On the actual day of testing, a few students within each grade were absent. These students were either tested on an alternate day, or replaced with a different randomly-selected student by the test administrator (second author of the study). Both the easyCBM© CCSS Math benchmark and the SAT-10 tests were administered online.

Measures

This study evaluated the relation between two assessments, the easyCBM© CCSS Math winter benchmark for Grades 6-8, and the Intermediate 3 (Grade 6), Advanced 1 (Grade 7), and Advanced 2 (Grade 8) versions of the SAT-10. All measures were administered online during the winter of 2014. Below, we discuss the technical adequacy evidence for the two assessments.

easyCBM© CCSS Math. The original development of the easyCBM© CCSS Math assessments is described in full by Anderson, Irvin, Patarapichayatham, et al. (2012). Broadly, the measures were developed with trained teachers serving as the primary item writers, who wrote all items to align with the Common Core State Standards (CCSS). These items were then piloted with students across the United States. The piloting plan included common items between grades so all items could be placed on a common, vertical scale. Items were evaluated primarily for their fit to the Rasch model expectations and difficulty. All items included a minimum of 200 responses. Items were then assembled into a set of 13 alternate forms within each grade, with the

average difficulty and distribution of item difficulties in each form matched. Each form contained 25 items. Three forms were designated for benchmarking, while the remaining 10 were reserved for progress monitoring.

In the first year of operational use, Anderson, Alonzo, and Tindal (2013b) examined the reliability of the measures and found the forms to be operating at less than ideal levels (i.e., Cronbach's alpha < .70). The forms were then revised during the summer of 2013 (Anderson et al., 2013a). An additional 5 items were included in each of the progress monitoring forms, and 20 additional items were included in the benchmark forms (for 30 and 45 item tests, respectively). The 5 additional items included were originally written to the National Council of Teachers of Mathematics (NCTM) Focal Point Standards, but had been judged to be adequately aligned with the CCSS in a formal alignment study (Irvin, Park, Alonzo, & Tindal, 2012). The additional 15 items included in the benchmark were linking items between forms, either vertically (common items between grades) or horizontally (common items between test forms within the same grade). Wray, Lai, Alonzo, and Tindal (2014) investigated the reliability of the CCSS Math tests using a large, extant dataset. The authors found Cronbach's ranged from .92 to .95 across Grades 6-8 for the fall and winter form during the 2013-14 school year. Split-half reliability ranged from .80 to .87 for the first half and .92 to .95 for the second half, while the correlation between the split-half forms ranged from .62 to .73. These results suggest strong internal consistency.

The current study is the first of multiple planned studies empirically investigating the validity of the measures (e.g., construct, predictive, etc.). Previous research has investigated the content validity of the items by exploring the match between the item stimulus and the corresponding CCSS the item was intended to measure, as judged by content experts (Anderson,

Irvin, Alonzo, & Tindal, 2012). Anderson et al.'s (2012) study suggested that, overall, the items had a high degree of alignment with the CCSS.

SAT-10. The full mathematics portion of the SAT-10 was administered to all students in this study. The Intermediate 3, Advanced 1, and Advanced 2 forms were administered to students in Grades 6-8, respectively, during the winter of 2014. The SAT-10 was designed to "encourage students to think and to enable them to demonstrate the extent to which their mathematics instructional programs have empowered them" (Pearson, 2004, p. 63). Each test form contained 80 multiple-choice items, with 48 addressing *Mathematics Problem Solving* and 32 addressing *Mathematics Procedures*. Test blueprints were based on recommendations from the National Council of Teachers of Mathematics (2000). All items were scored dichotomously (correct/incorrect) and scaled with a Rasch model. Students scoring at the 50th percentile had a scale score of 639, 655, and 663 for Grades 6-8, respectively. Total scale scores were used in this study.

The reliability of the SAT-10 has been assessed using a large extant dataset (*n* range = 1,824 to 3,484) with the Kuder-Richardson Formula 20 (KR 20), which is interpreted similarly to Cronbach's alpha (Pearson, 2004). KR-20 was estimated at 0.94 to 0.95 across grades, while the standard error of measurement (SEM) ranged from 3.93 to 3.95 items correct across grades. On average, students responded to approximately 37-43 items correct. Evidence for the validity of the SAT-10 stems primarily from the developmental process used. Following item piloting, all items were reviewed for bias by an advisory panel. All items were also reviewed by content experts. These reviews, combined with individual item statistics (e.g., Mantel-Haenszel, fit to the Rasch model), determined whether items were ultimately included in operational test forms.

Analyses

Correlation and regression analyses were conducted to explore the relation between the easyCBM© CCSS Math winter benchmark and the SAT-10. Correlation analyses were used to examine the raw, bivariate relation between the measures, while regression analyses were used to explore the degree to which students' performance on easyCBM© predicted their performance on the SAT-10 (i.e., the variance in SAT-10 accounted for by easyCBM©).

One of the benefits of regression analysis is that additional control variables can be added to the model so the unique relation between the measures can be more reliably evaluated. For example, previous research suggests that students who are enrolled in special education or English language learner programs, who are female, non-White, and/or eligible for free or reduced price lunch all generally perform lower on mathematics tests (Fryer & Levitt, 2009; Lee, 2002; Lubienski, 2001). These variables could be included in the model so the variance in SAT-10 uniquely accounted for by easyCBM© (net of demographic variables) could be evaluated. However, we anticipated having very low power to detect the effect of demographic variables given the relatively small overall sample size, and the even smaller number of students within the target group (e.g., the sample of special education students ranged from 6-8 students, across grades). We therefore included demographic control variables in exploratory models, but focus our presentation of the results primarily on the simple linear regression models.

Results

The bivariate correlations between the easyCBM© CCSS winter benchmark and the SAT-10 were .82, .77, and .75 for Grades 6-8, respectively. These relations are displayed in Figures 1-3, along with boxplots of each variable on the plot margins. Results from the simple linear regression analyses are displayed in Table 2. Overall, the easyCBM© CCSS measure accounted for 56% to 67% of the variance in the SAT-10, which was significant in all cases ($p <$

.001). For every one point increase in easyCBM©, students scored, on average, approximately 3.5-4.5 points higher on the SAT-10. Similarly, a one standard deviation increase on easyCBM© corresponded to, on average, approximately three-quarters of a standard deviation increase on the SAT-10, across grades.

The results of our exploratory multiple regression models, which included student demographic variables, are displayed in Tables 3-5. Across grades, no demographic variables were significant predictors of students' SAT-10 performance, with the exception of special education status in Grade 7. The small number of students represented within each group likely contributed to the observed effects, as the standard errors were large in all cases.

Discussion

The purpose of the easyCBM© assessment system is to provide teachers with information to help inform their educational decisions. If the tests do not adequately measure the skills they purport to measure, however, the validity of their use within decision-making contexts is reduced. The purpose of this study was to explore, preliminarily, the extent to which the easyCBM© CCSS Math tests measure the skills they purport to measure. The study can only be viewed as preliminary given that evidence for the validity of a test for an intended purpose is gathered through the accumulation of research over time, from multiple perspectives and with multiple methods. This study was the first empirical study investigating the validity of the easyCBM© CCSS Math tests. Across Grades 6-8, we explored the relation between the easyCBM© CCSS Math tests and the SAT-10, which has documented validity evidence (Pearson, 2004).

Overall, the results suggest that the two measures—easyCBM CCSS Math and SAT-10 Math—related highly with each other. In other words, if one knew the score a student received

on one test, one could predict the score on the other test with reasonable accuracy. Across grades, the measures correlated at 0.75 or higher. Further, the simple linear regression analyses accounted for 56%-67% of the total variance in SAT-10 scores. These results suggest that easyCBM© and the SAT-10 likely measure the same underlying construct, providing concurrent validity evidence for the use of the easyCBM© CCSS Math tests within RTI.

The overall modest sample size was perhaps the primary limitation to this study. While *a priori* power calculation suggested the sample size would be adequate for detecting statistical significance (and indeed it was), the generalizability of the sample is likely low. In other contexts, the magnitude of the relation between the easyCBM© CCSS Math measures and the SAT-10 may be lower or greater than observed here. Further, power calculations were based on a balanced design. While not a primary purpose of the study, demographic variables were dramatically unbalanced, and power to detect these effects was low. The coefficients for the demographic variables in Tables 2-5 should, therefore, be interpreted with caution.

Future research should continue to examine the validity of the easyCBM© CCSS Math tests for both adequately measuring students' math skills (i.e., the purpose of this study) and for informing teachers' educational decisions. It is quite possible, for instance, that teachers are provided with accurate information, but misinterpret the data. While this study marks the first empirical investigation into the validity of the easyCBM© CCSS Math measures' use, no research to date has investigated easyCBM© data reporting mechanisms.

The validity of the measures' use within RTI, however, depends on both aspects. Further, no research to date has investigated students' growth on the revised easyCBM© CCSS Math tests (note that the study conducted by Anderson, Saven, Irvin, Alonzo, & Tindal, 2014, used the 2012-13 version of the test forms). Future research should investigate both average growth, and

growth by student subgroups (i.e., those represented in Table 1). This study represented one preliminary step to holistically evaluating the validity of the easyCBM© CCSS Math tests use within an RTI framework.

References

Anderson, D., Alonzo, J., & Tindal, G. (2013a). easyCBM CCSS math item scaling and test form revision (2012-2013): Grades 6-8 (technical report 1313). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2013b). Study of the reliability of CCSS-aligned math measures (2012 research version): Grades 6-8 (technical report 1312). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). The alignment of the easyCBM middle school mathematics CCSS measures to the Common Core State Standards (technical report 1208). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Irvin, P. S., Patarapichayatham, C., Alonzo, J., & Tindal, G. (2012). The development and scaling of the easyCBM CCSS middle school mathematics measures (technical report 1207). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Saven, J. L., Irvin, P. S., Alonzo, J., & Tindal, G. (2014). *Teacher Practices and Student Growth in Mathematics: Grades 6-8* (Technical Report 1401). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Fryer, R. G., & Levitt, S. D. (2009). *An Empirical Analysis of the Gender Gap in Mathematics: Working Paper 15430 NBER Working Paper Series*. Cambridge, MA: National Bureau of Economic Research.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why and how valid is it? *Reading Research Quarterly, 41*, 93-99. doi: 10.1598/RRQ.41.1.4

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012). The alignment of the easyCBM grades 6-8 math measures to the Common Core Standards (technical report 1230). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31*, 3-12.

Lubienski, S. T. (2001). *A second look at mathematics achievement gaps: Intersections of race, class, and gender in NAEP data*. Paper presented at the Annual meeting of the American Educational Research Association, Seattle.

Messick, S. L. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi: 10.1037/0003-066X.50.9.741

National Council of Teachers of Mathematics. (2000). *Priniciples and standards for school mathematics*. Reston, VA: Author.

Pearson. (2004). *Stanford Achievement Test Series Tenth Edition: Technical Data Report*: Author.

Wray, K., Lai, C. F., Alonzo, J., & Tindal, G. (2014). *Internal Consitency and Split-Half Reliability of the easyCBM CCSS Math Measures, Grades K-8* (Technical Report No. 1405). . Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Table 1

Sample Demographics

Demographic variable	Grade		
	6 (*n* = 67)	7 (*n* = 63)	8 (*n* = 64)
Female	33 (49)	24 (38)	38 (59)
Special education (total)	8 (12)	7 (9)	6 (63)
Intellectual disability	1 (2)	3 (5)	0 (0)
Communication disorder	1 (2)	0 (0)	0 (0)
Other health impairment	1 (2)	1 (2)	1 (2)
Autism spectrum disorder	0 (0)	0 (0)	1 (2)
Learning disability	5 (7)	3 (5)	4 (6)
English Language Learner (total)	6 (9)	3 (5)	0 (0)
Declined services	0 (0)	1 (2)	0 (0)
ESL class period	1 (2)	0 (0)	0 (0)
LEP monitoring 1	4 (6)	1 (2)	0 (0)
LEP monitoring 2	0 (0)	1 (2)	0 (0)
No additional program	1 (2)	0 (0)	0 (0)
Lunch Eligibility (total)	42 (63)	36 (57)	29 (45)
Free	36 (54)	31 (49)	22 (34)
Reduced	6 (8)	5 (8)	7 (11)
Ethnicity			
American Indian or Alaskan Native	0 (0)	0 (0)	1 (2)
Asian	1 (2)	0 (0)	0 (0)
Black or African American	0 (0)	2 (3)	0 (0)
Hispanic	7 (10)	7 (11)	10 (16)
Other	3 (5)	5 (8)	2 (3)
White	56 (84)	49 (78)	51 (80)

Note. Proportions displayed in parentheses

ESL = English as a second language
LEP = Limited English proficient

Table 2

Simple Linear Regression Results

Grade	R^2	*F* (*df*)	Parameter	Estimate		95% CI	
				Standardized	Raw	Lower	Upper
6	0.67	129.4 (1, 65)	Intercept	-	568.21	549.07	587.35
			easyCBM©	0.82	3.95	3.26	4.65
7	0.59	87.86 (1, 61)	Intercept	-	596.57	575.81	617.34
			easyCBM©	0.77	3.53	2.78	4.28
8	0.56	78.19 (1,62)	Intercept	-	593.69	564.10	623.29
			easyCBM©	0.75	4.38	3.39	5.38

Note. All parameters significant, $p < .001$. easyCBM© represents the slope coefficient for the easyCBM© CCSS Math measure.

Table 3

Multiple Regression Results: Grade 6

Parameter	Estimate		95% CI		Sr^2
	Standardized	Raw	Lower	Upper	
Intercept	-	564. 53	537.81	591.25	-
FRL	-0.07	-5.34	-17.34	6.66	0.004
Female	0.14	10.46	-0.89	21.80	0.017
ELL	0.06	8.05	-13.42	29.53	0.003
SPED	0.02	2.67	-16.58	21.93	0.001
easyCBM©	0.82	3.98	3.18	4.79	0.494

Note. The overall model was significant, $F = 27.60$ (5, 61), $R^2 = 0.69$.

Table 4

Multiple Regression Results: Grade 7

Parameter	Estimate		95% CI		Sr^2
	Standardized	Raw	Lower	Upper	
Intercept	-	611.68	582.42	640.95	-
FRL	0.05	4.07	-9.66	17.80	0.002
Female	-0.06	-4.49	-17.57	8.58	0.003
ELL	-0.03	-5.74	-35.57	24.09	0.001
SPED	-0.24	-29.70	-52.26	-7.14	0.044
easyCBM©	0.67	3.07	2.18	3.95	0.305

Note. The overall model was significant, $F = 20.20$ (5, 57), $R^2 = 0.64$.

Table 5

Multiple Regression Results: Grade 8

Parameter	Estimate		95% CI		Sr^2
	Standardized	Raw	Lower	Upper	
Intercept	-	603.59	565.58	641.61	-
FRL	-0.11	-9.31	-23.72	5.10	0.012
Female	0.08	6.49	-8.13	21.11	0.006
SPED	-0.06	-9.02	-37.47	19.42	0.003
easyCBM©	0.70	4.09	2.92	5.25	0.355

Note. No students in the Grade 8 random sample were coded as receiving English language learner services. The overall model was significant, $F = 20.13$ (4,9), $R^2 = 0.\ 58$.

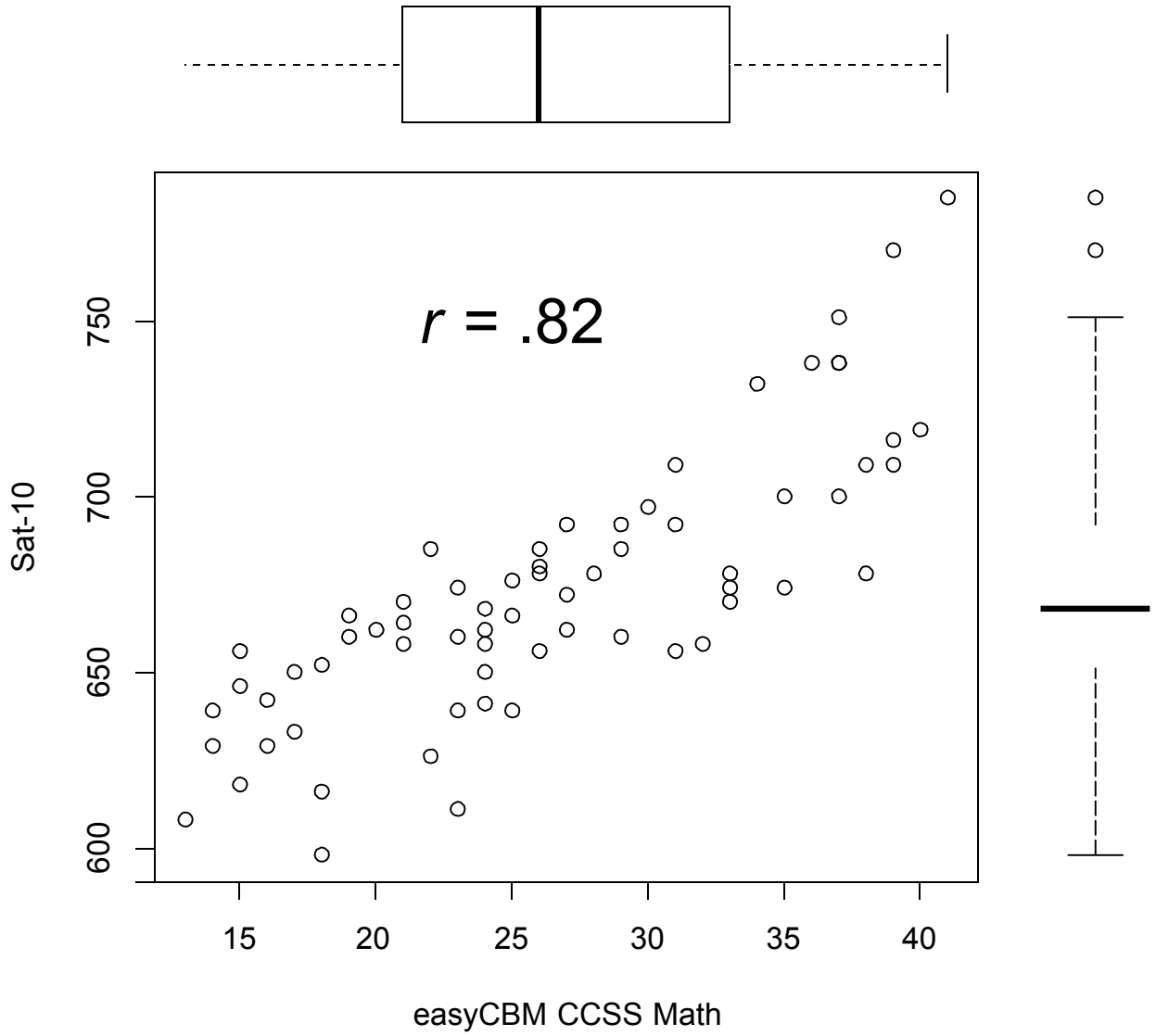

Figure 1. Grade 6 bivariate relation between easyCBM© CCSS Math winter benchmark and the SAT-10. Note that the univariate distributions for each variable are plotted on the margins

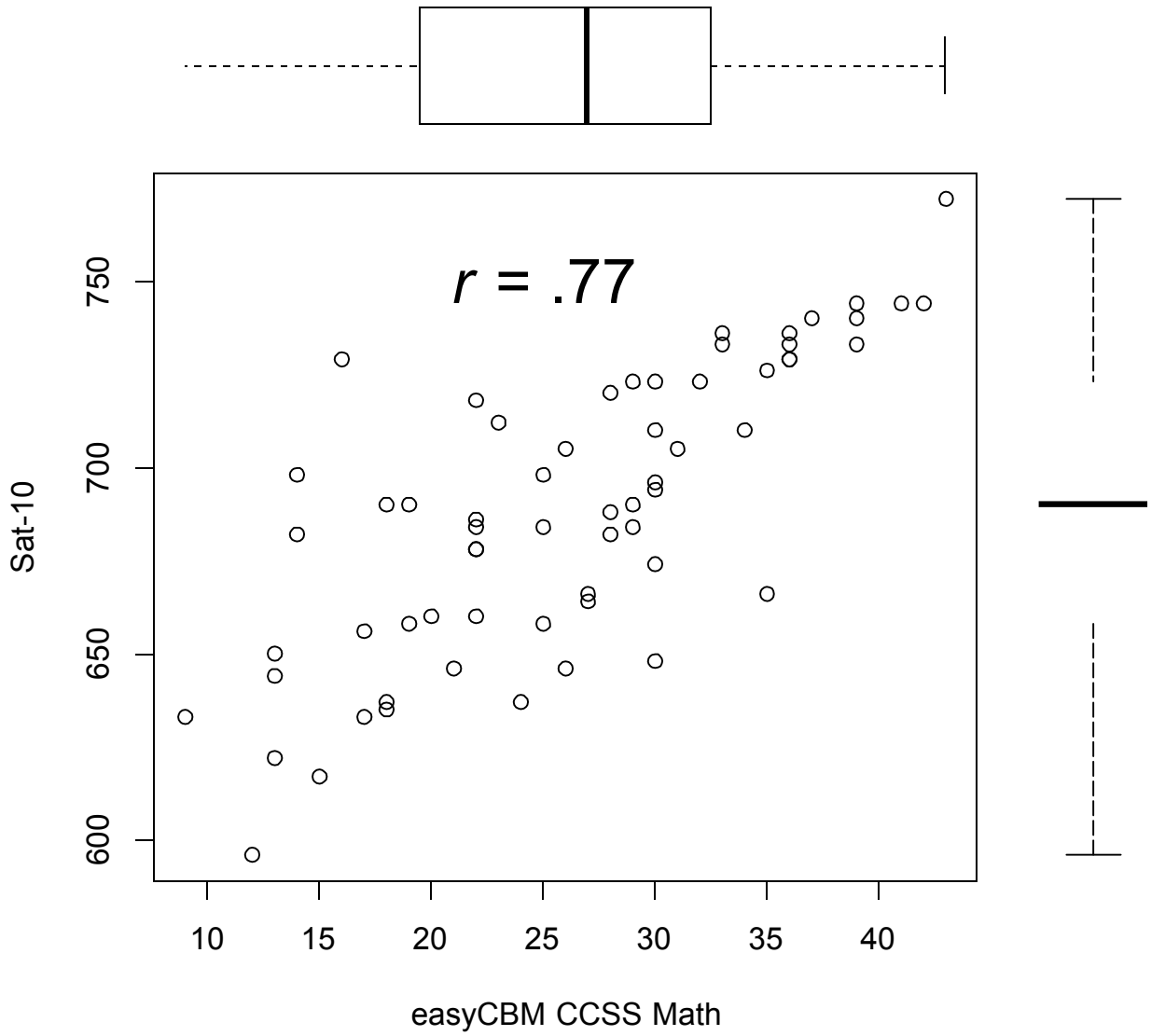

Figure 2. Grade 7 bivariate relation between easyCBM© CCSS Math winter benchmark and the SAT-10. Note that the univariate distributions for each variable are plotted on the margins

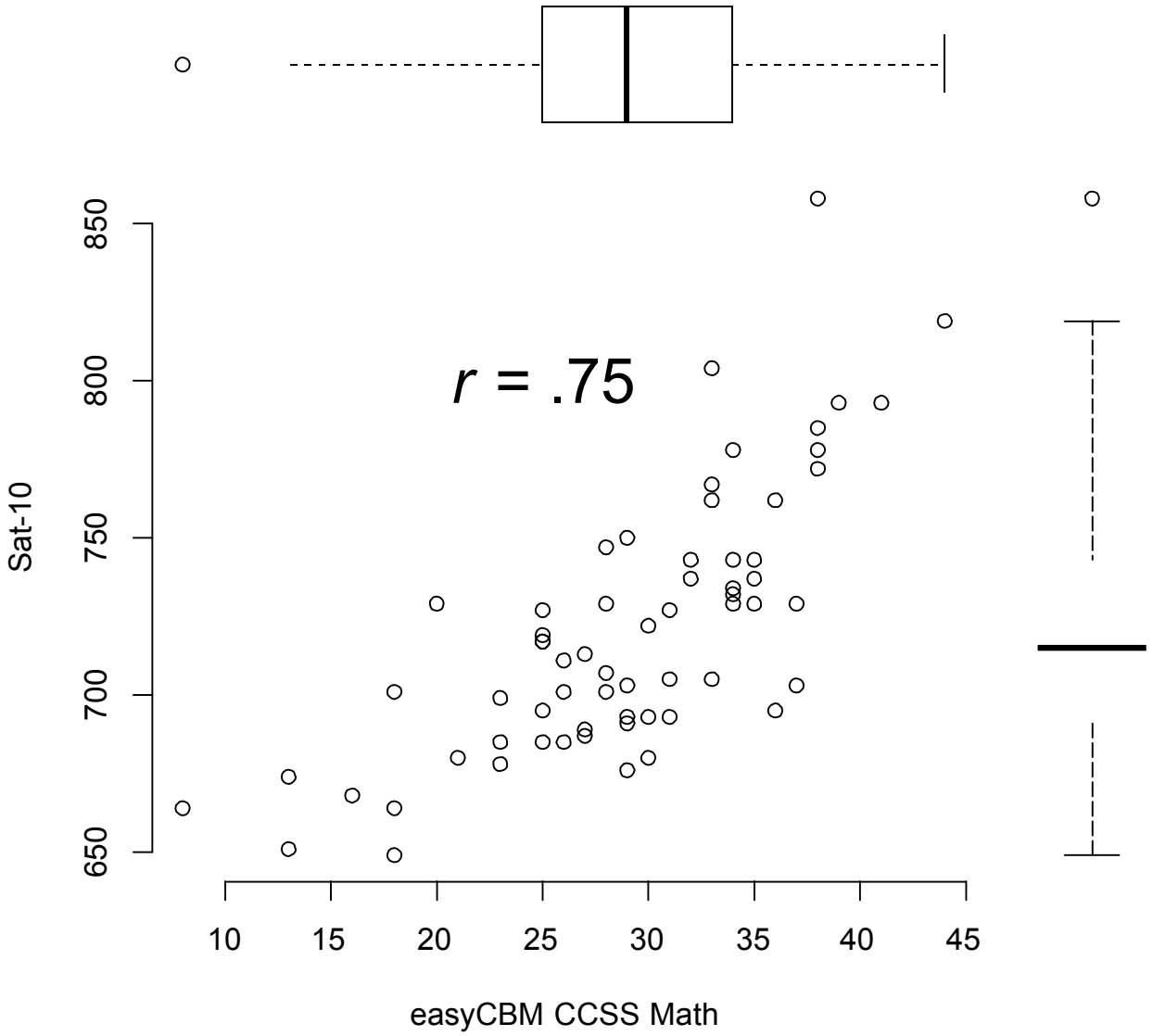

Figure 3. Grade 8 bivariate relation between easyCBM© CCSS Math winter benchmark and the SAT-10. Note that the univariate distributions for each variable are plotted on the margins