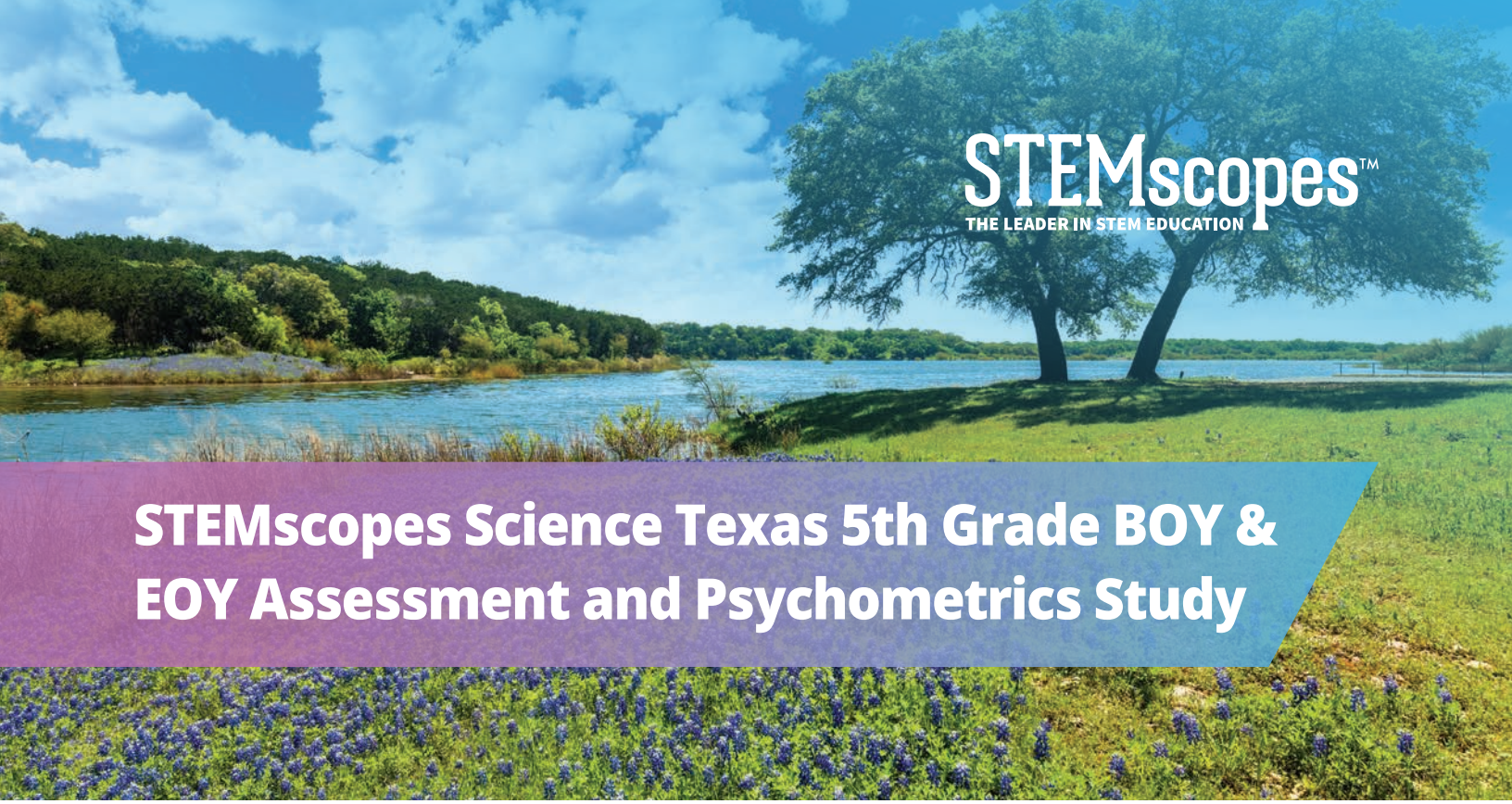
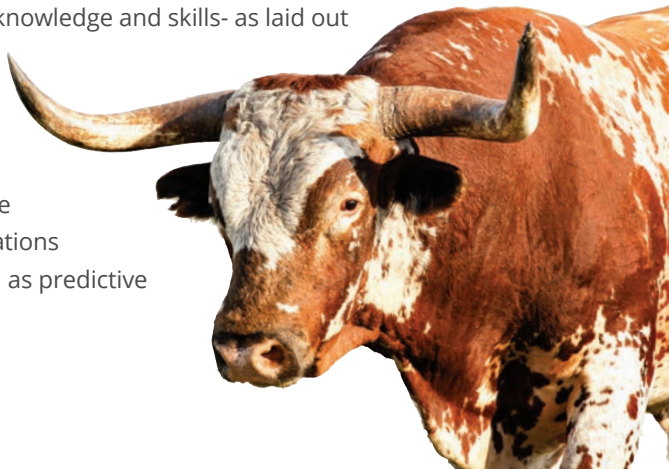


STEMscopes Science Texas 5th Grade BOY & EOY Assessment and Psychometrics Study

This study focuses on validating the STEMscopes Science 5th grade assessments. These are student-completed measures that are completed at the beginning and the end of the school year to evaluate science learning. The assessments are aligned to the Texas Essential Knowledge and Skills (TEKS) by grade level, and are designed to assist teachers in tracking student science learning progress. All items were written by Texas teachers (5th grade was written by teachers who were currently elementary teachers and had taught 5th grade for at least one year). All content writers were trained on how to analyze the Texas standards including verbs, content (nouns) and the context (e.g., did the standard include such as or including statements). That information was then used to create the appropriate questions to assess the standard. We had a peer review of the test items where other teachers reviewed the questions for alignment to standards and compared them to released questions from State of Texas Assessments of Academic Readiness (STAAR) Science assessment in order to verify STEMScopes assessment questions were at a similar level as the STAAR questions but unique material. After all questions were peer verified, they were reviewed by a content expert with a PhD for content accuracy. Once all items were approved, the BOY/EOY assessment blueprints were built to mirror the STAAR blueprints with a similar number of items to assess readiness and supporting standards. In 5th grade that means there were questions from 3rd, 4th and 5th grade standards. A curriculum specialist also verified that the questions on each test were testing similar content given some standards could have a wide range of content to test. This process provides initial validity regarding the assessment's content. This evidence is essential because it is this process that defines the general science content that will be appropriately assessed/measured. All later aspects of validity (such as how well the assessments relate to each other and to other science assessments) are related to this type of validity...that the test actually measures what it is supposed to measure: 5th grade science knowledge and skills- as laid out by 5th grade teachers in relation to the TEKS. The rest of the current report focuses on the statistical evaluation of the validity and reliability of the assessments based on a sample of 714 fifth grade students. We investigated each assessment's reliability and provided evidence that all items are measuring the same concept (general science) and that items are consistent. We also evaluated concurrent and predictive validity via correlations between the STEMscopes assessments and the STAAR assessment, as well as predictive models accounting for the unique nesting of students in classrooms.



PARTICIPANTS

We enrolled 5th grade students from New Braunfels ISD to evaluate the psychometric properties of the STEMscopes Science BOY and EOY tests. As reported to the Texas Education Agency for the 2021-2022 school year, the majority of students were White/Caucasian (48%), followed by Latin/Hispanic (45%), belonged to two or more races/ethnicities (4%) or were Black/ African American (2%). Approximately 1% were other races/ethnicities. Approximately 91% of students were native english speakers. District wide, approximately 36% of students were considered economically disadvantaged and 11.6% received special education services.

ANALYTIC STRATEGY

To evaluate the reliability and validity of the measure, several different types of analyses were employed. To investigate evidence of construct validity and reliability, we used confirmatory factor analysis with 1 factor specified. In other words, we expected the items to measure "general science knowledge" and little else. Once factor structure was confirmed, we evaluated individual items and tested two Item response theory (IRT) models to evaluate whether we should focus on item difficulty alone (1PL) or item difficulty and item discrimination (how well an item classifies an individual student's underlying knowledge). Once the factor and IRT models were complete, reliability and validity were also examined based on three key criteria:, internal consistency of all items, as well as predictive, and concurrent validity.

PRE-ANALYSIS DATA INSPECTION

The most well-known way to look at an item's difficulty is to calculate what percent of students answered that item correctly. Individual item percent correct on the 5th grade BOY ranged from 19% of students responded correctly to 89% of students responded correctly. The item percent correct at EOY ranged from 27% - 89%. This suggests the items cover a wide range of difficulties. The average number correct at BOY was 19.46 (SD = 6.21, min=0, max =32). The average total percent correct was 54% at BOY. Please note, no students received a perfect score of 36. At EOY the average number correct was 23.72 (SD = 6.31, min = 2, max = 36). The average total percent correct at EOY was 66%, and several students did achieve a perfect score.

FACTOR ANALYSIS

We performed a confirmatory factor analysis (CFA) on the 36 categorical items (i.e., correct versus incorrect responses) with separate models for BOY and EOY. Factor analysis is a statistical analysis that can be used to determine if all of the items are generally measuring the same thing (such as underlying science knowledge/ skills) and can be considered a test of construct validity (that it is measuring what it should) and reliability (that it is measuring consistently across items). Examination of model fit indices indicated that the one factor model fit the data very well at BOY (CFI = 0.98, RMSEA = 0.01), as well as EOY (CFI = 0.99, RMSEA = 0.01) given that CFIs greater than .95 and RMSEAs less than .05 are indicative of excellent fitting models (Hu & Bentler, 1999). This means there is evidence that the BOY and EOY test items, respectively, are, together, measuring the same underlying knowledge. Examination of the individual items indicated that two items on the BOY test were not functioning as hoped. Specifically, item 30 and item 19 did not correlate highly with the overall "general science factor." At EOY, only one item was not performing well: item 22. This means these items may not be giving teachers very much information about a student's knowledge and skills related to science.

Confirmatory factor models, like the one in the model mentioned above, could also be considered the same as an IRT 2 parameter (2PL) model where items are characterized by both their difficulties and their discrimination (how well the classify between students with different science knowledge and skills). Specifically, how highly the item is related with the underlying "general science knowledge" tells us how good that item is at discriminating among student's science ability. Items with a number close to or higher than 1.00 are better at measuring "general science" because they are more strongly related

to said knowledge. In the model described above the discriminations for BOY ranged from -.12 (this was item 30 and why we consider it a poor item as it is NOT describing science knowledge) to 1.19. At EOY, they ranged from 0.13 (item 22) to 1.18. Both tests include quite a bit of range in how well items discriminate among student science knowledge and skills.

As a follow-up to the above models, we ran an additional model for both BOY and EOY called an IRT 1 parameter model. In this model, all items are constrained to have the same discrimination. In the case of the current model, the constrained item discrimination at BOY was 0.82, and 0.91 at EOY. Thus, the only thing that differs between items in these models is their difficulty (please recall: when we calculate student's percent correct scores, we assume a model where only item difficulty differs and item discrimination is the same, but by comparing the IRT 1PL to the IRT 2PL, we can test this assumption directly). In the case of the current study, the IRT 2PL model fits better than the IRT 1PL for BOY and EOY.

At this point, we removed items 30 and 19 from the BOY test and item 22 from EOY. The modified IRT 1PL and 2PL models fit better for both BOY and EOY than the original models with all items, but the IRT 2PL still had slightly better statistical fit. However, it is not uncommon with larger sample sizes for a 2PL model to fit better, but there is a question of whether the estimation of the discrimination "is worth it" compared to the ease of a 1PL/"percent correct" model that anyone can calculate a student score for (i.e., number correct/ total items). One way to evaluate whether one should keep the 2PL model is to run a correlation between the outputted model scores based on the 2PL model (that is scores that take into account item difficulties and discriminations), and percent correct scores. If they are highly correlated, then one can feel confident maintaining the easier to understand "percent correct" (1PL) model that only accounts for item difficulty. In the case of the current study, the correlation between student's percent correct scores and the 2PL model outputted scores was 0.98 for both BOY and EOY indicating the scores are nearly identical. With this in mind, we keep the easier to use and easier to understand "percent correct" model.

ALPHA RELIABILITY

Next as an additional reliability check, we calculated Alpha reliability for the 36 items, as well as the 34 and 35 item modified BOY and EOY tests. This statistical analysis is used to provide additional evidence that the items are measuring the same thing (i.e., internal consistency of the items), with scores closer to 1.00 indicating higher levels of reliability. Alpha reliability for the 36 item BOY test was 0.82 and 0.84 for EOY indicating good reliability based on field standards. As expected, removing items 30 and 19 from the BOY test resulted in even higher reliability with alpha = 0.84, and the EOY test without item 22 had an alpha reliability of 0.85. Based on these analyses on the STEMscopes Science BOY and EOY 5th grade assessments, we conclude that the items are reliably measuring the same underlying "science knowledge and skills" as expected for each test respectively, and that providing teachers with student percent correct scores gives them the information they need to understand students' underlying level of science knowledge and skills.

In the following **Table 1** we present each item with its corresponding item difficulty. For ease of interpretation, we use the item percent correct scores to indicate difficulty and color codes such that difficult items (ranging from 0.00 - 0.49) are coded in red, moderately difficult items (ranging from 0.50 - 0.74) are coded in green, and easy items (ranging from 0.75 - 1.00) are coded in blue.

After that, in **Table 2**, we present the aggregate test characteristics (e.g., Mean, Median, Mode) for both the BOY and EOY tests. As can be seen in the two tables, the beginning of year test is quite a bit more difficult for this sample of students than the end of year test, potentially signaling student growth in science knowledge and skills.

TABLE 1 ITEM CHARACTERISTICS

ITEM #	TEKS covered BOY	Difficulty BOY	# respondents BOY	TEKS covered EOY	Difficulty EOY	# respondents BOY
1	TX 5.7B	0.33	630	TX 5.5A	0.71	666
2	TX 4.7C	0.73	630	TX 5.9A	0.88	666
3	TX 3.6B	0.75	630	TX 5.6C	0.73	666
4	TX 5.5A	0.63	630	TX 3.10B	0.73	666
5	TX 5.7A	0.47	630	TX 5.9C	0.45	666
6	TX 5.5B	0.47	630	TX 5.10B	0.49	666
7	TX 3.10B	0.64	630	TX 5.8B	0.59	666
8	TX 5.8A	0.39	630	TX 5.7B	0.75	666
9	TX 5.6B	0.68	630	TX 5.6A	0.89	666
10	TX 5.7A	0.63	630	TX 3.5C	0.84	666
11	TX 5.10A	0.37	630	TX 5.6B	0.52	666
12	TX 5.6C	0.33	630	TX 5.10A	0.71	666
13	TX 4.8C	0.37	630	TX 5.7A	0.69	666
14	TX 5.6A	0.66	630	TX 5.8C	0.7	666
15	TX 5.6D	0.37	630	TX 4.7A	0.88	666
16	TX 5.5A	0.52	630	TX 3.9A	0.76	666
17	TX 5.8B	0.5	630	TX 5.6C	0.68	666
18	TX 5.6B	0.8	630	TX 5.10A	0.56	666
19	TX 5.9A	0.19	630	TX 5.9B	0.58	666
20	TX 5.8C	0.81	630	TX 5.6A	0.85	666
21	TX 3.5C	0.79	630	TX 5.7A	0.6	666
22	TX 5.10B	0.62	630	TX 5.5B	0.27	666
23	TX 5.9B	0.69	630	TX 4.8C	0.29	666
24	TX 3.9A	0.76	630	TX 5.7B	0.67	666
25	TX 5.6A	0.66	630	TX 5.8A	0.7	666
26	TX 5.9A	0.45	630	TX 5.9A	0.84	666
27	TX 5.9C	0.33	630	TX 5.6D	0.71	666
28	TX 5.10B	0.67	630	TX 5.5A	0.83	666
29	TX 5.10A	0.89	630	TX 4.7C	0.76	666
30	TX 5.6C	0.29	630	TX 5.10B	0.69	666
31	TX 3.8D	0.38	630	TX 5.9B	0.42	666
32	TX 5.7B	0.29	630	TX 3.6B	0.75	666
33	TX 4.7A	0.57	630	TX 3.8D	0.69	666
34	TX 5.9B	0.53	630	TX 4.8B	0.62	666
35	TX 4.8B	0.37	630	TX 5.5A	0.36	666
36	TX 5.5A	0.54	630	TX 5.6B	0.55	666

TABLE 2 **TEST CHARACTERISTICS**

	BOY	EOY
Total Questions	36	36
Mean	19.46	23.72
Median	20	25
Mode	23	28
Mean Difficulty	54%	66%
Reliability	0.82	0.84

CONCURRENT AND PREDICTIVE VALIDITY

To evaluate concurrent and predictive validity, we ran correlations between the STEMscopes BOY and EOY assessments, and the State of Texas Assessments of Academic Readiness (STAAR) standardized science assessment. Next, we ran multilevel regression models to predict student STAAR science outcomes based on the STEMscopes Science Assessments. Multilevel models account for the fact that students are clustered by teachers (that is we would expect that students who were taught by the same teacher may have more similar scores than students taught by a different teacher). We ran two separate models: the first with BOY predicting STAAR (predictive validity because the tests occurred ~8 months apart), then EOY predicting STAAR (concurrent validity as they occur close to each other in time, about 2 weeks apart). At this point, we used the modified assessments (without the three items mentioned above). Results indicated strong correlations between the BOY and EOY STEMscopes science assessments ($r = 0.68$, $p < .001$), between BOY and STAAR ($r = 0.67$, $p < .001$), and between EOY and STAAR ($r = 0.73$, $p < .001$). Similarly, the multi-level models suggest strong associations such that a 1-point increase on the BOY STEMscopes science assessment is estimated to be associated with a significant 60.43 point gain on the STAAR science assessment scale score ($p < .001$). Likewise, a 1 - point increase on the EOY STEMscopes assessment is associated with a significant 64.77 point gain in the STAAR science scale score ($p < .001$). Overall, this provides strong evidence in support of the concurrent and predictive validity of the STEMscopes science assessments.

CONCLUSION

In summary, the BOY and EOY STEMscopes 5th grade science assessments conform to the measurement of general science knowledge and abilities as laid out by the TEKS. In addition, scores from the measure are internally consistent, and validity has been established with scores being strongly correlated with STAAR science scores. The STEMscopes science assessments provide teachers with easy-to-administer assessments with automatic grading via the STEMscopes web platform. These assessments can assist teachers with individualizing classroom instruction and tracking student progress.