Responses to Claims Raised by Walter Stroup

Briefing Document

This document provides responses to claims made by Dr. Walter Stroup before the House Committee on Public Education and to various media sources in the past few months. Dr. Stroup's remarks about Item Response Theory (IRT) and the variance between year-to-year test scores indicating solely a student's "test-taking ability" represent unpublished, non-peer-reviewed conclusions he has drawn from work he conducted or supervised previously. Dr. Stroup's controversial and unsubstantiated claims go against more than 50 years of scientific research in the field of testing and measurement and will not stand up to a technical review by qualified educational measurement experts.

August 20, 2012

Responses to Claims Raised by Walter Stroup

Briefing Document

Executive Summary

In June of 2012, Dr. Walter Stroup appeared before the Texas House Committee of Public Education and made misleading claims about the Texas statewide assessments drawing on studies that have not been vetted or validated through an academic peer review process for published research. To address these claims, experts from Pearson's assessment team serving the Texas program have authored this brief both to expose the serious flaws in Dr. Stroup's conclusions and to discuss the significant strengths of Texas's system of standards and assessments. This brief makes the following points:

- Item Response Theory (IRT), a widely recognized and validated statistical method, is critically significant in addressing two major concerns to fairness in testing. IRT allows for inclusion of test questions with varying difficulty levels while assigning student test scores fairly and identifies test questions that might be culturally biased. IRT takes into account difficulty in much the same way Olympic judges take into account the difficulty of a particular gymnastic routine or a swimmer’s dive in the calculation of a score.

- IRT does not rank order students or select test questions. IRT simply measures students’ academic knowledge and skills on a scale (like a ruler) and, just as a child gets taller, when students increase their knowledge and skills, their test scores will increase. IRT provides a thorough and fair measurement of growth and mastery.

- The development of the new Texas Essential Knowledge and Skills (TEKS) approved by the State Board of Education and the State of Texas Assessments of Academic Readiness (STAAR) tests that are aligned to those standards involved Texas educators in all phases. Thousands of Texas educators participated in more than 250 meetings to review test questions and set performance standards that fit today's needs. The tests are developed in a rigorous, scientific, and transparent way in order to produce scores that are fair, valid, and reliable. National experts, the Texas Technical Advisory

Committee, and the U.S. Department of Education routinely conduct independent reviews as part of the test development process.

- STAAR, and the previous assessment system, Texas Assessment of Knowledge and Skills (TAKS), are sensitive to student gains in proficiency that result from improved classroom instruction.

- A moderate correlation between consecutive grades of TAKS or STAAR tests is natural in that the standards are explicitly designed to reflect the progression of learning from grade to grade.

- Performance on these tests depends on how well a student mastered the pre-requisite knowledge and skills in prior years as well as how well the student learned to standards in the current year's instruction.

- Students will not score well by demonstrating only what they learned the previous year or in a different subject. The standards for each content area are linked grade to grade in a stair-step approach. To perform well on a test, a student will typically demonstrate proficiency on foundational content learned in previous grades and instruction in the current year.

The new STAAR tests are designed to better capture student learning and growth to the new higher standards. Texas conducted years of validity research in order to support the development and implementation of an aligned system to prepare students to graduate ready for postsecondary success. As a result, Texas's strong standards and well-designed assessments can play an important and helpful role in ensuring young Texans can meet the challenges of college and career.

Claims made by Dr. Stroup

Walter Stroup is an Associate Professor at the University of Texas at Austin in the College of Education—Curriculum and Instruction Department.

In June of 2012, Dr. Stroup appeared before the House Committee of Public Education and made controversial and unsubstantiated claims about the state’s assessment program that go against more than 50 years of scientific research in the field of testing and measurement. More specifically, Dr. Stroup claimed, before the committee and to various media sources in the subsequent months, that:

1. Item Response Theory (IRT) produces “a glitch […] in the DNA of the state exams that […] suggests they are virtually useless at measuring the effects of classroom instruction.”

2. IRT produces an exam that is more sensitive to how it ranks students based on that model than to measuring any gains in their year-to-year learning.

3. 72% of the variance between state assessments over a two-year period is the result of “test-taking ability” (without the student “knowing” any content) as opposed to student learning.

Data pertaining to Dr. Stroup’s claims

The data used as the foundation for Dr. Stroup’s claims appears to have originated from two studies: Alexander and Stroup (2006) and Pham (2009).

In the first study, Alexander and Stroup (2006) studied the effectiveness of a mathematics intervention program administered to 79 students in Richardson ISD against a control group and the general student population.

Though Alexander and Stroup (2006) reported mixed findings related to statistical significance for the various models they tested, they concluded that the intervention program was a success based on the fact that the intervention group outperformed the control group.

While Alexander and Stroup (2006) do not make any specific claims about IRT or “test-taking ability” in their study, the data that resulted from this research seem to have been used by Dr. Stroup in formulating his most recent claims.

In the second study, Pham (2009) subjected student TAKS data to a computer simulation designed for modeling complex behavioral interactions in biological systems. Based on the performance in this computer model (which is unrelated to IRT) the researcher drew conclusions about the workings of IRT.

Alexander, C. & W. Stroup (2006). *Richardson Math Project Final Report: Math TAKS Results*. Austin, TX, University of Texas at Austin.

Pham, V. (2009). *Computer Modeling of the Instructionally Insensitive Nature of the Texas Assessment of Knowledge and Skills (TAKS) Exam*. Unpublished Dissertation. The University of Texas at Austin.

Research base for Dr. Stroup's latest claims

Dr. Stroup's most recent claims before the House Committee for Public Education, and to various media sources shortly thereafter, represent unpublished, non-peer-reviewed conclusions he has drawn from work he conducted or supervised previously.

The academic peer-review process exists primarily to prevent researchers from asserting wild and debatable claims about established research practices without first having those conclusions reviewed by qualified experts. While the process ensures that all results and conclusions are vetted, replicable, and available for refutation in a controlled academic environment, it is particularly necessary when contentious claims are being asserted.

At the House Committee for Public Education and in subsequent interviews provided to the media, Dr. Stroup has openly stated:

1. That his research was unpublished but is available on his website.
   House Committee on Public Education – 6/19/12

2. That he was planning on submitting his finding to numerous journals but had not done so.
   Texas Tribune – 7/30/12

3. That he initially delayed preparing the findings for academic journals because of assurances he received that the STAAR exams would address the issues his research pointed to in the TAKS.
   Texas Tribune – 8/9/12

4. That he regretted the fact that his work had not yet been published, and that this fact had distracted from an examination of its technical merit
   Texas Tribune – 8/9/12

> *While unsubstantiated, non-peer-reviewed claims play very well in the media, and generate a lot of momentum around a particular political cause, they have no place in a serious discussion about the role of state-wide standardized testing.*

An examination of the technical merits of a researcher's claims is supposed to occur during the peer-review process—not by general invitation at a public hearing. The process provides an initial filter to determine whether there is any merit to the research design, methods applied, and conclusions of a professor's work.

In addition, no one at either the Texas Education Agency (TEA) or Pearson (the state's vendor tasked with producing the assessments) made any assurances to Dr. Stroup about the future test design of the State of Texas Assessments of Academic Readiness (STAAR) tests based on his previous work.

Dr. Stroup's mischaracterization of his own findings

Even before responding to the general claims made by Dr. Stroup, it is worth noting that a cursory examination of the materials he has made available reveal that he has mischaracterized his own findings to both the House committee and the media—something the peer-review process would have flagged immediately if he had chosen to follow that route.

More specifically, one of Dr. Stroup's main claims—that 72% of the variance of recent test scores can be predicted from the previous year's test scores—was derived from his lack of understanding of the difference between a regression coefficient and variance.

It should be noted that the 72% value that has been widely cited in the media, and by Dr. Stroup himself, does not actually appear in his unpublished study.

Dr. Stroup's unpublished research details a regression coefficient of 0.759. He inaccurately draws the conclusion that this means that the prior year's test score accounts for 75.9% of the variance in the current year's test score—revealing a fundamental misunderstanding of basic statistics. In reality, the variance he observed was 50.4%.

Dr. Stroup then goes on to claim, without any evidence, that the variance he observed between year-to-year test scores is the result of "test-taking ability." This statement is impossible for Dr. Stroup to defend given the nature of his research and something that a panel of qualified experts would have rejected during a review of this work.

Although test scores from year-to-year are related to one another, this relationship is primarily a function of the mastery of prerequisite skills necessary for success in subsequent years. This relationship—between test scores across multiple years and their direct connection to prior learning—is discussed in detail in the following sections.

Response to Dr. Stroup

In response to Dr. Stroup's claim, this section answers the following questions:

1. What is Item Response Theory (IRT) and how does it work?
2. Is IRT a valid method for scoring students?
3. Does IRT limit instructional sensitivity by selecting questions to rank-order students?
4. What role do Texas educators play in the development of content standards, tests questions, and performance standards?
5. How are the tests developed?
6. Does the fact that test scores are related to one another from year to year mean that classroom instruction does not play a role in student test scores?
7. Does a correlation between test scores suggest students are not learning?
8. Can you provide an example of how a prerequisite skill is necessary for success in subsequent years?
9. Can students score well by using only what they learned the previous year?
10. Do the state tests measure "test-taking skills" instead of learning?
11. Are the STAAR tests sensitive to instruction in the classroom?

1. What is Item Response Theory (IRT) and how does it work?

IRT is a statistical method commonly used for:

> *...design of tests, test assembly, test scaling and calibration, construction of test item banks, investigations of test item bias, and other common procedures in the test development process. Measurement researchers, public school systems, the military, and several civilian branches of federal government as well, have endorsed and employed IRT. (Hambleton, Swaminathan, & Rogers, 1991, p. vii)*

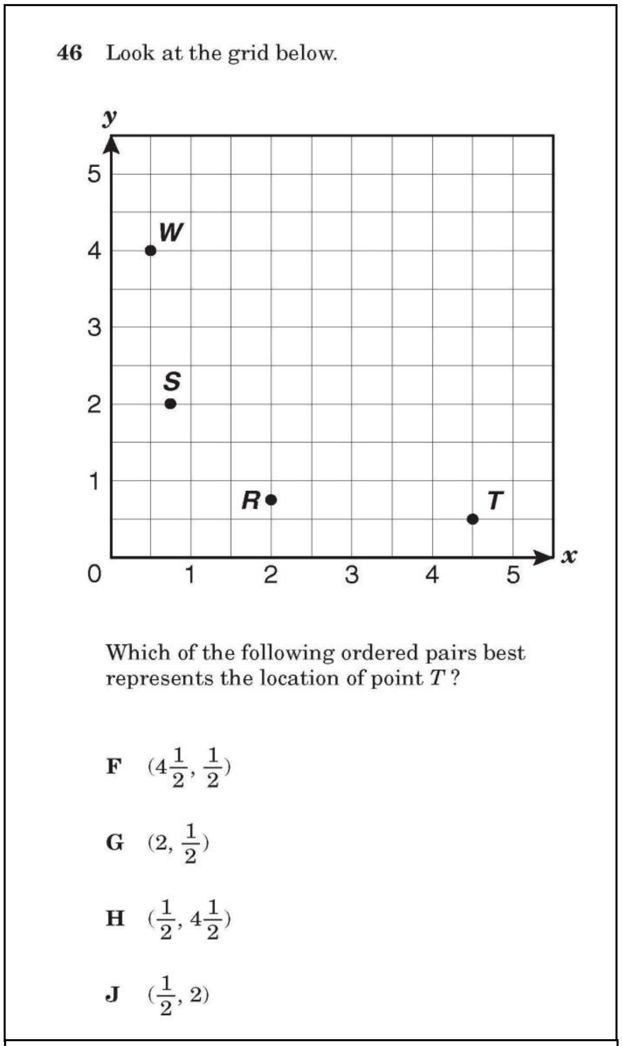
The primary body of research pertaining to IRT was established in the 1950s and 1960s and was widely adopted by the testing and measurement industry in the 1970s and 1980s after the availability of personal computers provided scientists the necessary means to perform IRT calculations.

IRT addresses two major concerns relative to fairness in testing. First, it allows for the inclusion of test questions with varying difficulty levels on an assessments while at the same time taking those varying difficulty levels into account when determining student test scores (illustrated below). Second, it allows the identification of test questions that might be culturally biased by examining the differential performance of student ethnic and gender groups on the test questions.

How does it work?

To show how IRT works, consider these questions from a released TAKS Grade 6 Mathematics assessment. The skill that each question assesses is shown below:

Easy	**Difficult**



46 Look at the grid below.

y
5
4
3
2
1
0
1
2
3
4
5
x
W
S
R
T

Which of the following ordered pairs best represents the location of point T?

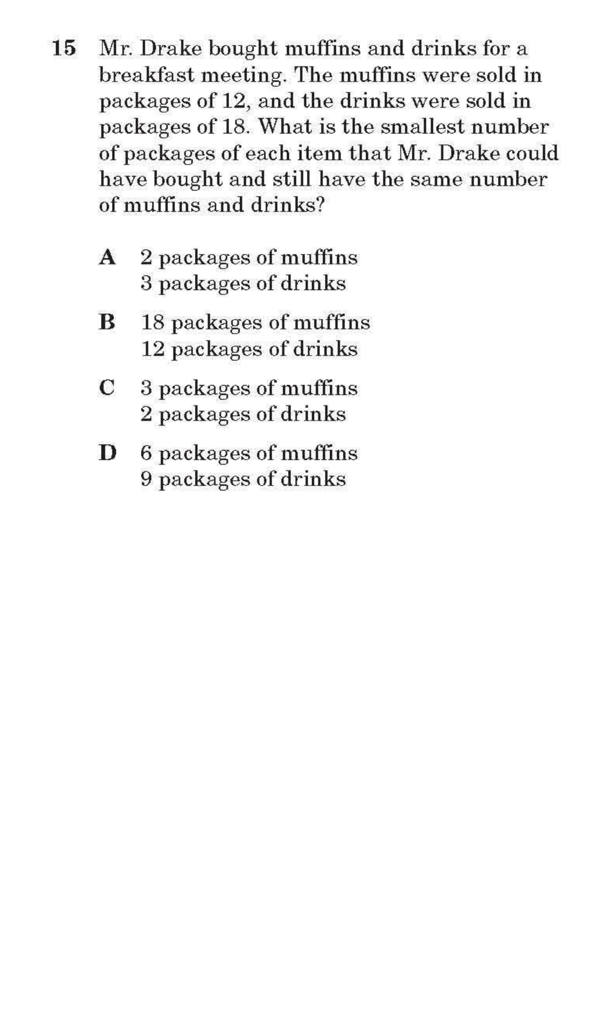
F $(4\frac{1}{2}, \frac{1}{2})$

G $(2, \frac{1}{2})$

H $(\frac{1}{2}, 4\frac{1}{2})$

J $(\frac{1}{2}, 2)$

Texas Essential Knowledge and Skill (TEKS)
Grade 6 Mathematics: Objective 3; 6.7A
Locate and name points on a coordinate plane using ordered pairs of non-negative rational numbers.



15 Mr. Drake bought muffins and drinks for a breakfast meeting. The muffins were sold in packages of 12, and the drinks were sold in packages of 18. What is the smallest number of packages of each item that Mr. Drake could have bought and still have the same number of muffins and drinks?

A 2 packages of muffins
3 packages of drinks

B 18 packages of muffins
12 packages of drinks

C 3 packages of muffins
2 packages of drinks

D 6 packages of muffins
9 packages of drinks

Texas Essential Knowledge and Skill (TEKS)
Grade 6 Mathematics: Objective 1; 6.1F
Identify multiples of a positive integer and common multiples and the least common multiple of a set of positive integers

The two previous questions show different levels of item difficulty within content standards from the same grade level. The easy question is cognitively simpler ("locate and name") and requires less math knowledge to answer correctly. The difficult question is cognitively more demanding ("identify") and requires more math knowledge in order to answer correctly. IRT lets us acknowledge these differences in difficulty before assigning students a test score.

Simply put, IRT lets us talk about student test scores in a way that takes into account the difficulty of questions, in much the same way that Olympic judges take into account the difficulty of a particular gymnastics routine.

Example

Pretend that Maria took a Grade 6 mathematics test in 2010 and her brother Tommy took a Grade 6 mathematics test a year later in 2011. If Maria's test was composed primarily of questions similar to the easy item above and Tommy's test was composed primarily of questions similar to the difficult item above, then, although both tests assessed similar content standards, Tommy's test would be significantly more difficult than Maria's.

Simply put, IRT gives us a way to talk about student test scores that takes into account the difficulty of questions.

However, without IRT, if both Maria and Tommy got 70% of the questions correct on their respective tests, one would simply conclude that they had similar overall knowledge and proficiency of Grade 6 mathematics concepts.

IRT allows us to acknowledge the varying difficulty levels of the questions on each test. Using IRT, Tommy's overall scale score on the assessment (his score adjusted for difficulty) would be much higher than Maria's because the questions on this test were more challenging and required more math knowledge.

IRT levels the field for evaluating academic performance regardless of the specific test questions a student takes and gives us a more powerful tool for understanding student academic proficiency than what than what can be obtained by simply looking at a percent-correct score.

Details on item difficulty and scale score conversion can be found in the Texas Technical Digest on TEA's website

2. Is IRT a valid method for scoring students?

Yes. IRT is a proven method for scoring students used worldwide.

IRT has been used for more than 40 years to support assessments such as the National Assessment of Educational Progress (NAEP), the Programme of International Student Assessment (PISA), the Graduate Records Exam (GRE), the Graduate Management Admission Test (GMAT), the Law School Admissions Test (LSAT), and state assessments developed to meet the requirements of No Child Left Behind.

Validation of the use of IRT in the literature:

> *"Although classical test theory (CTT) has served test development well over several decades, item response theory (IRT) has rapidly become mainstream as the theoretical basis for measurement (Embretson & Reise, 2000)."*
>
> *"Today, item response theory is being used by many of the large test publishers (Yen, 1983), state departments of education (Pandey & Carlson, 1983), and industrial and professional organizations (Guion & Ironson, 1983) to construct both norm-referenced and criterion-referenced tests, to investigate item bias, to equate tests, and to report test score information. In fact, the various applications have been so successful that discussions of item response theory have shifted from a consideration of their advantages and disadvantages in relation to classical test models to consideration of such matters as model selection, parameter estimation, and the determination of model-data fit (Hambleton, & Swaminathan, 1985)"*

3. Does IRT limit instructional sensitivity by selecting questions to rank-order students?

No. IRT does not enforce a rank-ordering of students across years nor does it select test questions. Test design and item selection are handled by content specialists and subject matter experts through a rigorous process that ensures the assessments measure the TEKS standards (described in the following sections).

IRT simply provides a measurement of students' academic knowledge and skills on a scale (like a ruler) where the students can grow and change over time. Just as when a child gets taller, his height increases, when students increase in their knowledge and skills, their test scores will increase. And just as with measuring height, some students will grow faster than others, and IRT will reflect these differences appropriately and accurately, providing a thorough and fair measurement of growth toward the mastery of the standards.

4. What role do Texas educators play in the development of content standards, test questions, and performance standards?

Content standards (the TEKS) are established by Texas State Board of Education (SBOE) through a multi-phased process that includes educators, parents, business and industry leaders, employers, and expert reviewers designated by the board. A complete revision of the TEKS across all content areas was initiated by the SBOE in the last several years and is still ongoing in some subjects.

The development process for both the TAKS and the STAAR involves Texas educators in all phases. These educators include K–12 classroom teachers, higher education representatives, curriculum specialists, and administrators. For STAAR, thousands of Texas educators participated in more than 250 meetings to review test questions and set performance standards for the assessments.

5. How are the tests developed?

The test development process is rigorous, scientific, and transparent.

TAKS and STAAR were developed through a process that adheres to the scientific methods defined in *The Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999). The tests developed by this process have been approved for technical adequacy through the federal peer-review process. Details about each phase of test development are disclosed in the annual technical digest published by the Texas Education Agency.

The test development process begins with the standards—the Texas Essential Knowledge and Skills (TEKS). These are the knowledge and skills taught in Texas classrooms. Subject matter experts (many of whom are current or former teachers) draft test questions to measure specific elements of the TEKS. These draft questions undergo multiple rounds of review involving additional subject matter experts and Texas educators. The review process verifies that the test questions are aligned to the standards and are accurate, reliable, and fair. For STAAR, thousands of Texas educators participated in more than 250 meetings to review test questions.

Potential test questions are then field-tested to evaluate how they performed and gather statistics for subject matter experts at TEA to review, evaluate, and ultimately make decisions about whether to use them on a future test.

Tests are then built so that the mix of questions represents the TEKS for the grade and subject, as indicated by the test blueprint. The test blueprint (published on the TEA website) defines the proportion of the test that is dedicated to each academic skill area. For example, skills that fall under "patterns, relationships, and algebraic reasoning" are measured by 8 of the 46 questions on the STAAR grade 3 mathematics test.

6. Does the fact that test scores are related to each one another from year to year mean that classroom instruction does not play a role in student test scores?

No.

There is a relationship between test scores from one year to the next. This relationship is called a correlation. The correlation depends on different factors such as the content being measured, how long the tests are, and how closely the content is aligned from one level to the next.

TAKS and STAAR are sensitive to student gains in proficiency that result from improved classroom instruction.

TAKS and STAAR are designed to assess performance on the academic standards (the TEKS), which are deliberately linked in order to promote the development of skills and knowledge from one grade to the next. Therefore, TAKS and STAAR show significant correlation between grades because they faithfully assess the linked standards.

Correlation doesn't mean that a student's score the next year can be predicted perfectly, and it doesn't mean that the test forces students into a rank order.

TAKS and STAAR are sensitive to student gains in proficiency that result from improved classroom instruction. Increased proficiency that results from instruction can be measured through increases in test scores.

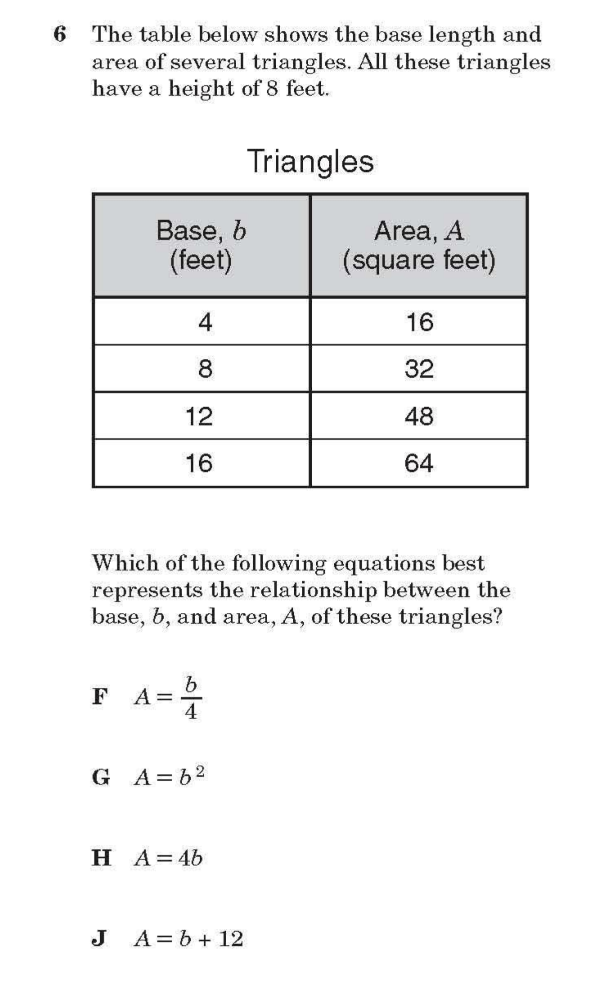
7. Does a correlation between test scores suggest students are not learning?

No. A moderate correlation between consecutive grades of TAKS or STAAR tests is appropriate.

This is because the TEKS define a progression of skills and knowledge that is aligned grade to grade. Student performance on the assessments across years depends to some degree on how well students master the prerequisite skills in the prior year's instruction as well as how well they build on those prerequisite skills in the current year of instruction.

8. Can you provide an example of how a prerequisite skill is necessary for success in subsequent years?

Student performance on TAKS and STAAR across years depends in part on how well students retain the skills acquired in prior years of instruction. This is how the tests are designed. TAKS and STAAR reflect that the standards for each subject become more demanding in each successive grade. Therefore, performance on the tests depends on how well a student mastered the prerequisite skill sets in the prior year's instruction as well as how well the student developed those skills in the current year of instruction. Consider these questions (the standard assessed by each is shown below):



6 The table below shows the base length and area of several triangles. All these triangles have a height of 8 feet.

Triangles

Base, b (feet)	Area, A (square feet)
4	16
8	32
12	48
16	64

Which of the following equations best represents the relationship between the base, b, and area, A, of these triangles?

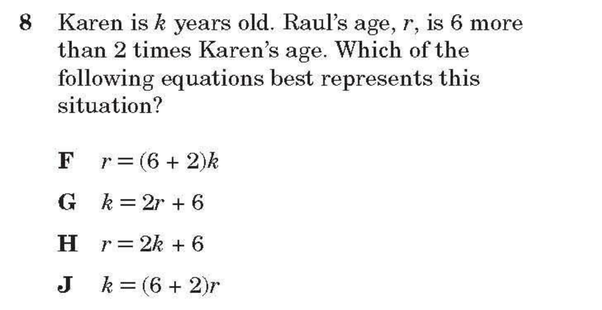
F $A = \frac{b}{4}$

G $A = b^2$

H $A = 4b$

J $A = b + 12$

Texas Essential Knowledge and Skill (TEKS)
Grade 6 Mathematics: Objective 2; 6.5A
Formulate equations from problem situations described by linear relationships.



8 Karen is k years old. Raul's age, r, is 6 more than 2 times Karen's age. Which of the following equations best represents this situation?

F $r = (6 + 2)k$

G $k = 2r + 6$

H $r = 2k + 6$

J $k = (6 + 2)r$

Texas Essential Knowledge and Skill (TEKS)
Grade 7 Mathematics: Objective 2; 7.5B
Formulate problem situations when given a simple equation and formulate an equation when given a problem situation.

The two questions show the progression of difficulty within the same skill family. The TAKS Grade 6 question is cognitively simpler (use model set data to formulate a simple equation). The TAKS Grade 7 question is cognitively more demanding (use a verbal description to formulate a complex equation). However, both questions require specific math knowledge about formulating equations to answer correctly.

The Grade 7 question builds on the skills in this area that students acquired at Grade 6. Understanding this concept in Grade 6 is a necessary foundation for learning how to extend the concept in Grade 7. However, without extending the skill through classroom instruction in Grade 7, the student would not necessarily be able to answer the Grade 7 question correctly.

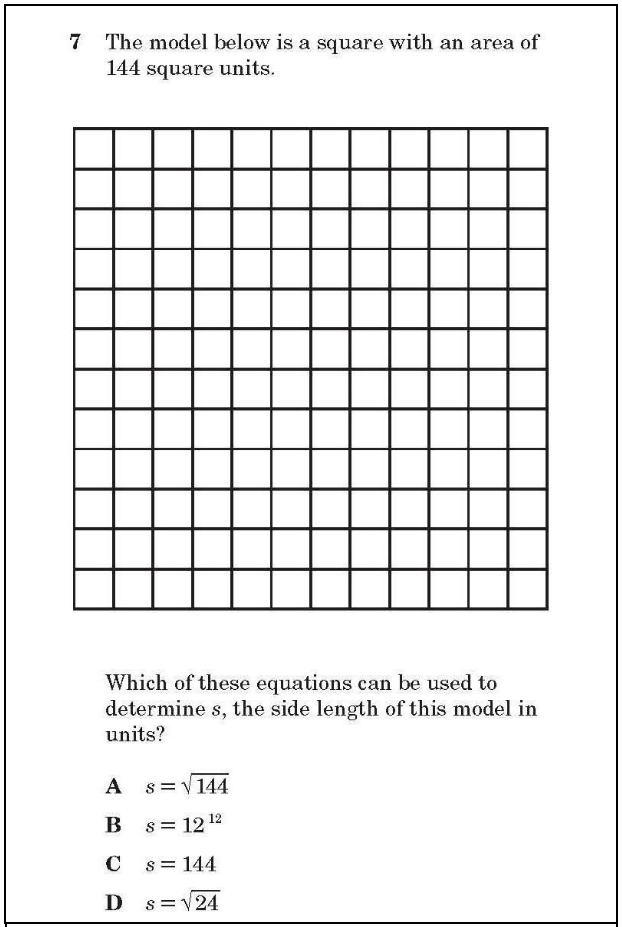
It is for these reasons that student performance on the tests across years depends to some degree on how well students have retained the skills acquired in prior years of instruction.

9. Can students score well by using only what they learned last year?

No. Each TAKS and STAAR test is designed to assess the state standards for a specific grade and subject.

The standards for each content area are linked grade to grade. To perform well on a test, a student cannot rely only on skills retained from previous years of instruction. The student must demonstrate proficiency in the standards for the current grade.

For example, the Numbers, Operations, and Quantitative Reasoning standard is defined and tested at Grade 6 and Grade 7, but the specific skill dealing with square roots is new in Grade 7. Without classroom instruction in Grade 7, students would not be expected to answer this question correctly:



7 The model below is a square with an area of 144 square units.

Which of these equations can be used to determine s, the side length of this model in units?

A $s = \sqrt{144}$

B $s = 12^{12}$

C $s = 144$

D $s = \sqrt{24}$

Texas Essential Knowledge and Skill (TEKS)
<u>Grade 7 Mathematics: Objective 1; 7.1C</u>
Represent squares and square roots using geometric models.

10. Do the state tests measure "test-taking skills" instead of learning?

No. Although test scores from year to year are correlated to one another, this correlation is a function of the mastery of prerequisite skills necessary for success in subsequent years.

11. Are the new STAAR tests sensitive to instruction in the classroom?

Instructional sensitivity is the degree to which student improvement in academic knowledge and skills—through better classroom instruction—can be measured via increases in test scores. Between 2003 and 2011, Texas students continually demonstrated achievement gains on the TAKS assessments. These increases reflected improvements in the instruction provided by schools focusing on the TEKS, the standards on which the Texas statewide assessments are based.

As student performance increased, however, the sensitivity of TAKS to continue to measure improvements in student performance was diminished because there was less and less room for students to demonstrate growth. In other words, by 2011, students had outgrown TAKS.

Consistent with a growing national consensus regarding the need to provide a more clearly articulated K–16 education program—one that focuses on fewer skills and addresses those skills in a deeper manner—the TEA began implementing a new assessment model for the STAAR tests at the elementary, middle, and high school levels.

Prior to the TEA's effort around implementing the new STAAR program, the Texas State Board of Education (SBOE) initiated a complete revision of the TEKS across all content areas (which is still ongoing in some subject areas).

The new assessment model for the STAAR program focuses on newly revised TEKS that are most critical to student success. This model better measures the academic performance of students as they progress from elementary to middle to high school.

In addition, the majority of the new STAAR assessments are designed to test content students study in their current year, as opposed to testing content studied over multiple years, strengthening the alignment between what is taught and what is tested.

The transition from TAKS to the more rigorous STAAR assessment was a necessary step in the evolution of the Texas Assessment Program, one that 1) allowed students to demonstrate growth, 2) focused on newly revised TEKS most critical to student success, and 3) articulated a clear connection between K–12 learning and college readiness.

As instructional improvement occurs in Texas schools moving forward, STAAR scores will reflect increases in student performance.