FR-08-22

HumRRO
Human Resources Research Organization

An Investigation of Illinois State Assessment System High School Scoring Patterns

Arthur A. Thacker
Emily R. Dickinson

Human Resources Research Organization (HumRRO)
10503 Timberwood Circle, Suite 101
Louisville, KY 40223
Phone (502) 339-9331
FAX (502) 339-9432

Prepared for: Illinois State Board of Education
100 N. 1st Street
Springfield, IL 62777

February 2008

Executive Summary

This study investigates the origins of an apparent divergence of WorkKeys Reading, a component of the Prairie State Achievement Examination (PSAE), from other components of the assessment. The issue was noted by a group of schools and districts concerned with the scoring patterns of the 2006 and 2007 cohorts of students. They noticed that students in 2007 scored much lower on WorkKeys Reading than the 2006 cohort, while 2007 students scored higher on all other components of the PSAE. The group was particularly concerned by an increase in ACT Reading, another PSAE component. ACT produces concordance tables linking ACT Reading with WorkKeys Reading, making the divergence seem unlikely.

The full population of student-level data was provided by ACT and the Illinois State Board of Education (approximately 130,000 cases per year). As a first step, the divergence was verified at the state level. This step confirmed that schools and districts noting the divergence had not mistakenly interpreted their data. It also helped ensure that the divergence was not isolated to only a few schools or districts. Finally, because the federal No Child Left Behind (NCLB) legislation (2002) requires that schools report percent proficient, it was necessary to verify that the divergence represented a mean shift in student scores rather than a statistical artifact created by the calculation of percent proficient. All of these potential explanations for the divergence were quickly discarded by simply calculating the means and standard deviations for each PSAE component test. WorkKeys Reading means, for the state student population, did indeed decline, while all other components improved from 2006 to 2007.

The magnitude of the decline for WorkKeys Reading was calculated using effect size statistics (Cohen's D) (Cohen, 1988) in order to compare the change scores across components. WorkKeys Reading declined by approximately 0.15 standard deviations from 2006 to 2007. The other PSAE components improvements ranged from 0.01 to 0.06 standard deviations. Compared to other states' assessments that are also used as part of NCLB, the decline in WorkKeys Reading scores for a single year represents a dramatic and anomalous change.

The next step in investigating the decline was to determine if one part of the WorkKeys Reading scale was more or less impacted than the other. This can occur if the test items change and no longer represent each part of the scale in the same way. For instance, if fewer items strongly discriminate toward the upper end of the scale, it is possible that fewer scale score points will be associated with raw scores toward that end of the scale. This could result in a reduction of the precision of the test to measure students in a particular part of the scale. That, in turn, could cause the transformation from raw score to scale score to result in some levels having fewer associated scale score points and might reduce the number of students at a particular level. If this were one of the higher levels, it might help explain why the overall mean declined. To investigate if this phenomenon or any similar issue that might impact only part of the scale, the percentage of students scoring at each level were compared for 2006 and 2007. The number of students scoring in the top three levels of WorkKeys declined by about 10% from 2006 to 2007 and there was no indication that any level was impacted more than another.

To further investigate the nature of the decline in WorkKeys Reading across the entire scale, a box-and-whiskers plot of each score level compared to its corresponding ACT score was created. The median ACT Reading scale score associated with each WorkKeys level was 1 to 2

points higher in 2007 than in 2006 for every possible level. By comparison, only one median ACT Math score plotted against WorkKeys Math level changed at all from 2006 to 2007, and that change was only a single scale score point. These analyses indicate that WorkKeys Reading scores declined for all levels of students in 2007 compared to 2006 in relation to ACT scores.

School-level means were created by aggregating student-level data. Analyses at the school level mirror the findings from the student level. At the school level, correlation tables relating scores from 2006 and 2007 were created. Those correlations indicated that schools' relative rank order from 2006 was largely maintained for 2007. This held true for WorkKeys reading as well as all other PSAE components. WorkKeys correlations for 2006 were somewhat attenuated because scale scores were not available. The extent to which level correlations were attenuated in 2006 can be estimated by examining level and scale score correlations in 2007. All results from the correlation study were within expected ranges.

Next, schools were matched by school codes included in the data sets to investigate the correlations across years. The correlations between like components of PSAE from 2006 to 2007 were all within expected ranges and very similar to each other. Similarly, change scores were calculated for each component and correlated with other components. Despite the decline in WorkKeys Reading, the correlations for all component change scores were very similar. This indicates that schools that gained on one component were likely to gain on all others in terms of overall rank order. In the case of WorkKeys Reading, schools that declined less were likely to gain on other components, whereas schools that declined more were likely to have gained less or declined on other components.

Finally, regression analyses and comparisons of effect sizes for student subgroups according to ethnicity, gender, economic status, and disability status were investigated. These analyses were conducted to ensure that the decline on WorkKeys Reading had not impacted any particular student subgroup more than another. At the school level, regression analyses suggested a potential bias for WorkKeys Reading for economically disadvantaged or disabled students. This pattern was not found at the student level statewide by comparing effect size differences for these subgroups. While the results indicate substantial achievement gaps among subgroups, there was no consistent finding to indicate a bias associated with the change in scoring pattern.

This study shows that the WorkKeys Reading decline was substantial and inconsistent with score patterns for all other PSAE components. The correlations indicate that change scores for all components were fairly consistent across schools. These results show that WorkKeys Reading was more difficult for 2007 students than for 2006 students relative to the other PSAE components. The extent to which changes in curriculum or student preparation might have impacted these patterns is unknown, but it seems unlikely that the entire state deemphasized the skills and knowledge required to score well on WorkKeys, to the advantage of the other components, within a single academic year. This is particularly troublesome given how closely related WorkKeys Reading and ACT Reading are expected to be. If the results from Illinois are similar to the national sample, it may be necessary for ACT to construct new concordance tables. If the results from Illinois are anomalous compared to the national sample, further investigation of the scoring, scaling, and equating processes may be indicated. While this study verified and quantified the decline in WorkKeys Reading scores compared to other PSAE components, no clear indication of the reason(s) for the anomaly was discovered.

One potential explanation for the PSAE scoring pattern observed between 2006 and 2007 relates to the equating of earlier administrations of the WorkKeys assessment. In 2004 and 2005, WorkKeys Reading was equated using a random equivalent groups design. We do not know a great deal more about the design, however, including how the groups were selected or created. Establishing equivalent groups is vital for ensuring equating stability using this method. WorkKeys Reading means for Illinois increased by a substantial margin from 2004 to 2005. In fact the scores increased by almost exactly the same amount that they decreased from 2006 to 2007 (based on estimating effect size from the mean shift from 2004 to 2005 and current variance data). When asked about this increase and the equating methodology, ACT informed HumRRO that the 2007 WorkKeys Reading assessment was equated to the 2004 administration, while the 2006 assessment was equated to the 2005 administration using a common-item non-equivalent groups design (where some items are repeated from an earlier administration and equating is accomplished using the change in scoring patterns from one administration to the next). Since the 2004 and 2005 assessments were subjected to an equating procedure, it should not have mattered which earlier administration was used for equating 2006 and 2007. However, without further information about the 2004/2005 equating procedure we can not know how well that procedure worked. It is possible that some methodological or random error might have contributed significantly to the apparent gain from 2004 to 2005. Then, when the 2006 administration is equated to 2005 and the 2007 administration to 2004 using common item equating, that same methodological or random equating error would result in an apparent decrease in mean scores. HumRRO has no means of investigating the likelihood of this possibility from the data provided, but the overall data patterns indicate that this is a plausible explanation for the decline.

An Investigation of Illinois State Assessment System High School Scoring Patterns

Table of Contents

Index of Tables

Index of Figures

Introduction

In Illinois, all eligible Grade 11 public school students take a battery of tests including the ACT English, Mathematics, Reading, Science and Writing tests, the WorkKeys Applied Mathematics and Reading for Information tests, and an internally developed science test (which assesses components of the state science standards not included on ACT). These components together make up the Prairie State Achievement Examination (PSAE). This study examines the relationships between scores on each of the components across the 2006 and 2007 administrations.

This study was conceived in part because of concerns that student scores on the ACT and WorkKeys components of the PSAE, especially the reading components, have diverged. Some Illinois schools and districts have noticed gains on one reading component and either no gains or even declines on the other (personal communication, Joyce Zurkowski, January 21, 2008). While this phenomenon is possible, it would be unlikely for such a pattern to be widespread across an entire state. Certainly, the test items on the two components are written to assess different aspects of the overall reading construct. It is possible that, through instruction or because of other factors, students might acquire the knowledge and skills required by one component at a different rate than they acquire the knowledge and skills required by the other. However, scores on these two components are typically very highly correlated. One would expect gains on one component to mirror gains on the other. ACT includes crosswalk information (e.g., a score of 5 on the WorkKeys is roughly equivalent to an ACT score of 22) in its reports on the alignment studies between the two components. Such information would not be appropriate if the two components were not highly related.

There are at least four potential reasons that schools and districts might report concerns about the divergence of the reading components:

1. The schools and districts reporting the divergence are anomalies and the overall state scoring pattern is as expected.
2. The schools and districts reporting the divergence have developed an erroneous perception and the state scoring pattern is as expected.
3. The schools and districts reporting the divergence are correct and the divergence is related to the content and/or design of WorkKeys compared to ACT.
4. The schools and districts reporting the divergence are correct and the divergence is related to score processing or equating of either of the component tests.

This study was designed to ascertain the extent to which the divergence in reading scores was genuine across all students and schools and to estimate the magnitude of the divergence.

Description of Data

The 2007 data for this study were provided by ACT on compact disk. The data file contained all pertinent scale scores, level scores (for WorkKeys data) and student demographic information. ACT also provided 2006 data, but the provided file was obviously incomplete (about 6,000 cases). The Illinois State Board of Education (ISBE) provided a student data file via

a secure File Transfer Protocol (FTP) site. The 2006 data was the most complete file available to ISBE and contained data on roughly 130,000 students. This was very similar to the case count for the 2007 file (133,597 students in the 2006 file, 134,074 students in the 2007 file). The 2006 data did not contain scale scores for the WorkKeys tests, however. Only level score was included on the WorkKeys file. Writing results were provided only for 2007.

All data were provided in text (flat file) format, organized such that each row of the file represented one student, with columns containing data keyed to a file layout document. SAS programs were written, using file layout documents provided by ACT, to convert the data to SAS data sets. The SAS data sets were then read into SPSS to facilitate analyses. Analyses performed at the school level used a data set aggregated by school code. The aggregated files contained school-level mean scores for each component test comprising PSAE. Analyses at the school level across years used a merged data set containing school-level results from 2006 and 2007 merged by school code.

Data Analysis

Student-level Score Divergence

It is certainly possible that the schools and districts with the greatest concern were anomalous in their data patterns. Without some indication of the number of schools experiencing the divergence in scores compared to the total number of schools in the state it is impossible to gauge how likely this explanation is to be true. In any event, this explanation may not be particularly satisfying to the affected schools. By computing student-level means, we can determine if the divergence reported by schools is a statewide phenomenon. Table 1 contains means and standard deviations from 2006 and 2007 across all Illinois students with reported scores. Table 1 also contains the difference (2007 mean – 2006 mean) for each reported PSAE component score. Finally, Table 1 contains effect size statistics (Cohen's D) for difference scores. The effect size statistics are in standard deviation units to allow for direct comparison among the various PSAE components (Cohen, 1988).

There are also several reasons schools and districts might have developed an incorrect perception that reading scores have diverged. The most obvious reason is the conversion of the two components to an overall reading score that is then reported as evidence of Adequate Yearly Progress (AYP) under the federal No Child Left Behind (NCLB) Act of 2001. It is possible that a school could make significant gains on one or the other component (or even both) and still not gain in terms of school-level percent proficient. It is even possible that the percent proficient could decline while one or the other or both component means increased. Because so much attention is given to NCLB, it might seem inconsistent for reading scores to be increasing while schools continued to fail to meet AYP targets.

If this were the case, a simple examination of the shift in mean scores at the student level should follow the expected pattern (no divergence). If the school-level divergence was not mirrored by the student-level mean shift, we could conclude that the anomaly might rest in the conversion of student scale scores to the NCLB reporting scale (percent proficient). However, Table 1 indicates that student-level results show a divergence of WorkKeys Reading scores with all other indicators of student performance. We must therefore conclude that the divergence

reported by schools and districts was not caused by a conversion anomaly, but legitimately represents the state population results.

Table 1. Student-Level Means and Effect Size Statistics

	Mean (06)	SD (06)	Mean (07)	SD (07)	Gain (07-06)	Effect Size Cohen's D
WorkKeys Math (Level)	4.58	1.746	4.63	1.772	0.05	0.029
WorkKeys Reading (Level)	4.68	1.494	4.46	1.404	-0.22	-0.152
ACT English (SS)	19.33	6.27	19.51	6.382	0.18	0.028
ACT Math (SS)	19.83	5.335	20.17	5.495	0.34	0.063
ACT Reading (SS)	19.75	6.322	19.82	6.115	0.07	0.011
ACT Science (SS)	19.67	5.151	19.85	5.104	0.18	0.035
ACT Composite (SS)	19.77	5.288	19.97	5.271	0.20	0.038

WorkKeys Reading scores dropped by approximately 0.15 standard deviations between 2006 and 2007. All other components of the PSAE increased. The increases ranged from 0.011 to 0.063 standard deviations. To put these gain scores in perspective, a recent study conducted by the Center on Education Policy (CEP) (2007) reported annual effect size changes for all states for which data were available from 1999 through 2006. While only 22 states had sufficient comparable data available (means and standard deviations) to allow for comparisons of effect sizes for some number of years, about 80 high school-level state effect sizes were calculated for the studied time period. Some states had only two years' worth of comparable data (enough to compute a single effect size), while others had complete data from 1999 through 2006. The calculated annual high school-level effect sizes from the CEP study are summarized in Table 2. The data in Table 2 show that PSAE components, other than WorkKeys Reading, were very similar to average annual gain scores in other states. The negative WorkKeys Reading change falls within the range of observed annual effect sizes in other states, but only 3 annual changes in other states, from 1999-2006, were larger than the 0.15 standard deviation decline observed in Illinois from 2006 to 2007. So, while not outside the range of national results, this decline would certainly be large enough to be considered an outlier.

Table 2. Data from CEP Study (2007) Depicting Comparison States' Annual Gains as Effect Size (1999-2006)

Subjects	Mean Effect Size	Effect Size SD	Range	Number of Cases
Math	0.0490	0.094	-0.147 to 0.379	80
Reading	0.0378	0.099	-0.293 to 0.342	81

Table 1 shows that nearly all mean scores for components of PSAE are improving over time. Table 2 demonstrates that this pattern of small annual improvement in mean scores is similar among other state NCLB assessments. It is common and expected for test scores to improve over time. This pattern was noted and made infamous by the "Lake Woebegone" papers of the 1980s, which hypothesized that cheating was largely responsible for the observed gains (Cannell, 1987; Cannell, 1989). Linn (1998) later described several legitimate (and less inflammatory) reasons why we should expect test scores to increase over time (e.g., students may become increasingly familiar with the format of the tests, repeated use of the same test for several years, changes in participation rates, focusing of instruction, etc.). The general tendency of test scores to increase makes the large decline in WorkKeys Reading scores even more puzzling.

ACT publishes national percentages of students scoring at each level. The published percentages do not differentiate individual WorkKeys components. Table 3 presents the published national averages (based on approximately 540,000 respondents) as well as Illinois results for both math and reading for 2006 and 2007.

Table 3. National Percentage Results for WorkKeys Compared to Illinois Data

WorkKeys Level	National Sample	Illinois Math 2006	Illinois Math 2007	Illinois Reading 2006	Illinois Reading 2007
7	7	9.6	12.5	6.8	4.0
6	18	22.4	22.3	20.9	15.3
5	34	29.3	23.5	34.1	32.4
4	45	16.2	20.0	25.1	34.7
3	17	15.2	14.7	7.9	8.6
Below 3	9	7.4	7.1	5.1	5.0

Table 3 shows that there is considerable variability between 2006 and 2007 for both WorkKeys Reading and Math statewide percentages at each level. The Illinois results are similar to the national sample. There was a precipitous drop in the percentage of students scoring in the top three categories for reading from 2006 to 2007. In 2006, 61.8% of Illinois students scored 5, 6, or 7 compared to 51.7% in 2007; or about 10% fewer students total. The percentage of students in the top three categories for math dropped from 2006 to 2007 as well, but only by 3%.

Figure 1 presents WorkKeys Math levels compared to ACT Math scale scores in box-and-whiskers plots. Each WorkKeys level has two boxes, the first for 2006 and the second for 2007. The center lines in the boxes represent the median ACT score for students at a particular WorkKeys level. The box represents the middle 50% of the students scoring at that WorkKeys level. The whiskers extending from the box indicate the range (without extreme outliers) of ACT scores at that level. Figure 2 contains the same information for WorkKeys Reading.

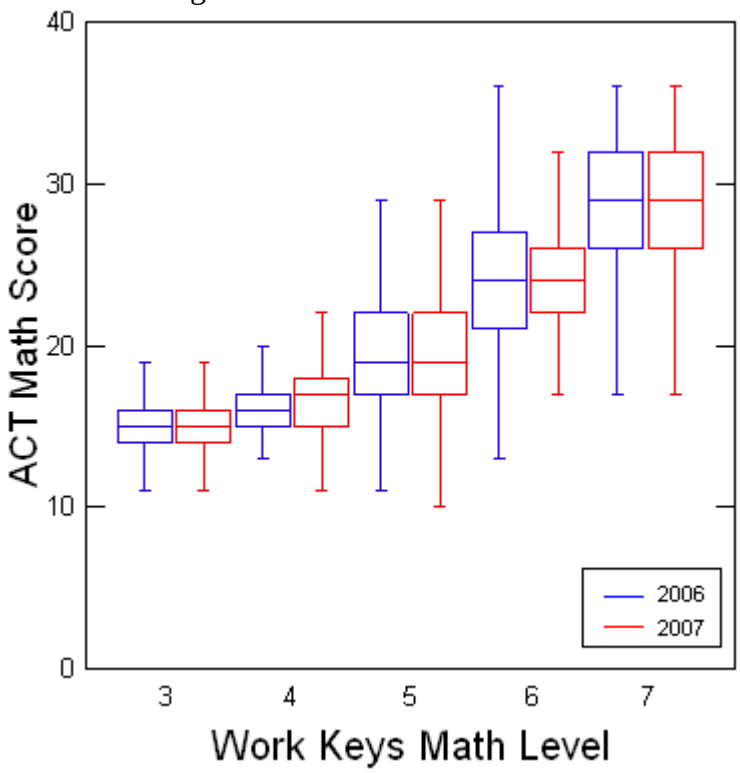

Figure 1. WorkKeys Math Level by ACT Math Scale Score box-and-whiskers plot.

From Figure 1, it is clear that, for math, the median ACT score for students within any particular WorkKeys level is very consistent from 2006 to 2007. The median lines in the centers of the boxes are identical except for a single point shift for Level 4. In addition, the boxes themselves (representing the center 50% of the data) show that the variance within each level has also remained fairly constant. Level 4 has a slightly higher range in 2007 and Level 6 has a slightly smaller range, but overall the results are very similar.

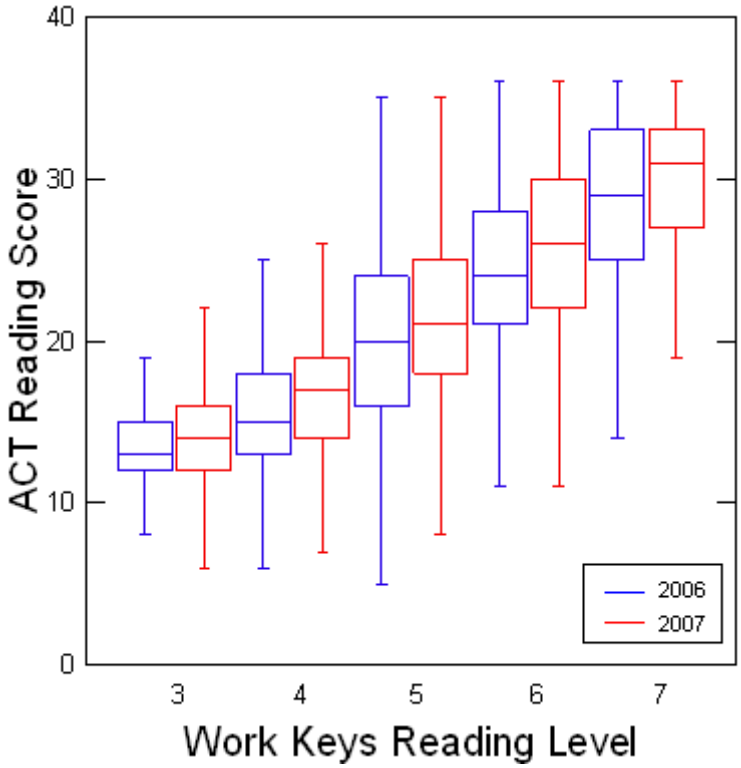

Figure 2. WorkKeys Reading Level by ACT Reading Scale Score box-and-whiskers plot.

Unlike Figure 1, Figure 2 shows a great deal of difference from 2006 to 2007 in reading. Particularly disturbing are the large and consistent shifts in median ACT score by WorkKeys level. For each level, the median corresponding ACT score was 1 to 2 points higher in 2007 than in 2006. A Level 3 would have most closely corresponded to an ACT score of 13 in 2006, but 14 in 2007. A Level 4 that would have corresponded with an ACT score of 15 in 2006 would correspond to 17 in 2007. The remaining levels follow the same pattern, indicating that WorkKeys Reading levels in 2007 correspond to higher ACT scores than in 2006. It is unknown if this pattern holds for the national sample or is unique to Illinois.

Student Level Correlations

In addition to examining means and effect size statistics, it is also informative to investigate the strength of the relationships between all of the component scores with each other. Prior studies have shown strong positive correlations among all of the ACT components (Bacci, Koger, Hoffman, & Thacker, 2003; Hoffman & Tannen, 1998). Those studies also reported a strong correlation between ACT scores and other test scores, and a somewhat weaker positive relationship between the ACT components and students' grades. Simply put, students who perform well on one component of ACT tend to perform well on all of them and vice versa. Similarly, students with higher scores on other indicators of achievement tend to score better on ACT than do students with lower scores. All of these correlations were positive and significant ($p < 0.01$) for Illinois students, as well. No further discussion of statistical significance (p values) is included in this report. The discussion is focused on the interpretation of the results rather than statistical significance.

The same correlation methodology was used for WorkKeys levels as for ACT scale scores. WorkKeys scale scores were not available for 2006, so in order to compare across administrations, we were forced to use WorkKeys level. For 2007, we can correlate WorkKeys scale scores with level scores and ACT scale scores as an indication of how much attenuation we might expect from using level instead. For both math and reading, the correlations are higher than 0.93, so while we would certainly have preferred scale scores, level provides a reasonable comparison across years.

Table 4 presents student-level correlations relating all PSAE components with one another for 2006 and 2007. We would expect similar content areas to yield the highest correlations. Then because of possible similarities in test-taking circumstances or other method effects, the next highest correlations are expected to be between different content areas within the same test. Finally, the lowest correlations are expected to be between different content areas on different tests. Specific test results on PSAE components (and ACT and WorkKeys components, as well) are often combined to create other reporting scales. These combination scores (e.g. ACT Composite, Reading Combined Scale Score) are also included in the correlation tables. Great caution should be used in interpreting combined score correlations since in many cases they will be artificially inflated because they are often correlated with their own constituent parts. Correlations presented in Table 4 are formatted as follows:

- Correlations between similar content areas across different tests (e.g., ACT Math with WorkKeys Math) (These correlations are in bold and underlined).
- Different content areas within the same test (e.g., ACT English with ACT Science) (These correlations are in italics).
- Different content areas within different tests (e.g., ACT Math with WorkKeys Reading) (These correlations are in bold, but not underlined).
- The relevant correlations from the 2006 study are included in parentheses immediately below the 2007 results for comparison purposes.
- Combined score correlations are presented for completeness but are not discussed.

For each correlation table, tests, or combinations of tests for which scores are reported, are listed in the first column and numbered sequentially. The remaining columns are each labeled with the corresponding test numbers. To find a particular correlation, simply find the tests of interest in the left-most column and the corresponding numbers of those tests. Then find the cell for which the numbers intersect to find the correlation. For example, if one were interested in the correlation of WorkKeys Reading Level (#4) with ACT Reading (#8), that correlation can be found where the column labeled #4 intersects with the row labeled #8. The two numbers in that cell (0.630 and 0.615) are the student-level correlations for those tests. The top number in the cell is the result for 2007, the bottom number in the cell (in parentheses) is the result for 2006. All subsequent correlation tables are arranged in the same manner.

Table 4. Student-level Correlation Table

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Work Keys														
1. Math Scale Score	1.00													
2. Math Level	.944													
3. Reading Scale Score	*.693*	*.665*												
4. Reading Level	*.643*	.633 (.641)	.939											
Internally Developed														
5. Science Scale Score	**.766**	**.720 (.702)**	**.693**	**.645 (625)**										
ACT														
6. English	**.740**	**.691 (.676)**	**.714**	**.656 (.659)**	**.746 (.726)**									
7. Math	**.838**	**.752 (.722)**	**.648**	**.585 (.580)**	**.758 (.745)**	*.785 (.782)*								
8. Reading	**.688**	**.633 (.615)**	**.695**	**.630 (.615)**	**.730 (.714)**	*.821 (.824)*	*.718 (.733)*							
9. Science	**.743**	**.681 (.678)**	**.671**	**.613 (.619)**	**.745 (.740)**	*.770 (.783)*	*.797 (.807)*	.751 (.770)						
10. Composite	**.822**	**.753 (.731)**	**.749**	**.682 (.676)**	**.815 (.797)**	*.932 (.930)*	*.899 (.899)*	*.907 (.915)*	*.900 (.909)*					
11.English_Writing	**.735**	**.696**	**.715**	**.664**	**.734**	.967	*.773*	*.804*	*.764*	*.912*				
Combined Scale Score														
12. Reading	.750	.709 (.706)	.927	.868 (.874)	.770 (.751)	.829 (.832)	.736 (.740)	.905 (.895)	.767 (.781)	.891 (.891)	.822			
13. Math	.961	.911 (.909)	.706	.654 (.666)	.795 (.786)	.792 (.789)	.937 (.924)	.726 (.728)	.796 (.804)	.887 (.880)	.788	.777 (.786)		
14. Science	.811	.763 (.753)	.732	.680 (.677)	.926 (.926)	.813 (.811)	.825 (.827)	.789 (.793)	.928 (.927)	.915 (.912)	.807	.826 (.828)	.852 (.858)	
15. Writing	.689	.661	.672	.632	.679	.876	.717	.742	.714	.840	.967	.768	.737	.755

Table 4 indicates that the convergent validity coefficients (similar subject correlations across tests) are within expected ranges. Similar to prior studies (Bacci, et al., 2003; Hoffman & Tannen, 1998) the 2007 ACT Math test correlates more highly with WorkKeys Math (0.838), than ACT Reading correlates with WorkKeys Reading (0.695) or ACT Science correlates with the Illinois Science test (0.745). It is common for mathematics tests to be more highly related than tests in other subjects.

Table 4 also shows that, as expected, the correlations using WorkKeys level were somewhat attenuated compared to correlations using scale scores. The 2007 ACT Math to WorkKeys Math correlation dropped from 0.838 to 0.752 and the ACT Reading to WorkKeys Reading correlation dropped from 0.695 to 0.630 when using level instead of scale score. Using level, however, allows comparison across years. The 2006 results in Table 4 (just below the 2007 results in each cell, in parentheses), were very similar in all cases to the 2007 results. The relationships between the relative rank-orderings of students from one test score to the other do not appear have changed appreciably, despite the decline in WorkKeys Reading scores.

School-Level Correlations

Similar to the student-level correlations, Table 5 depicts school-level correlations. The same formatting conventions are used. School-level data were calculated by simply aggregating (averaging) student scores for each indicated school code. It is typical for correlations of aggregated data to be larger than individual-level correlations, and this is the case for Illinois school-level correlations. Again, the 2007 data are presented in each cell of the table with corresponding 2006 data (where possible) included below (in parentheses) in the same cell.

Table 5. Correlations of 2007 School-Level Mean Scores With Corresponding 2006 Values in Parentheses

Variable	1	2	3	4	5	6	7	8	9	10	11	12
WorkKeys												
1. Math Scale Score	*1.00*											
2. Math Level Score	0.99	*1.00*										
3. Reading Scale Score	*0.91*	*0.90*	*1.00*									
4. Reading Level Score	*0.90*	*0.90*	0.99	*1.00*								
		(.86)										
Internally Developed												
5. Science Scale Score	**0.94**	**0.94**	**0.91**	**0.89**	*1.00*							
		(.93)		**(.85)**								
ACT												
6. English	**0.90**	**0.88**	**0.90**	**0.88**	**0.90**	*1.00*						
		(.86)		**(.83)**	**(.88)**							
7. Math	**0.94**	**0.91**	**0.87**	**0.84**	**0.91**	*0.92*	*1.00*					
		(.89)		**(.79)**	**(.90)**	*(.92)*						
8. Reading	**0.92**	**0.91**	**0.91**	**0.90**	**0.94**	*0.96*	*0.93*	*1.00*				
		(.88)		**(.84)**	**(.92)**	*(.96)*	*(.94)*					
9. Science	**0.92**	**0.90**	**0.90**	**0.88**	**0.93**	*0.95*	*0.95*	*0.96*	*1.00*			
		(.89)		**(.84)**	**(.92)**	*(.95)*	*(.95)*	*(.96)*				
10. Composite	**0.94**	**0.92**	**0.91**	**0.90**	**0.94**	*0.98*	*0.97*	*0.98*	*0.98*	*1.00*		
		(.90)		**(.84)**	**(.92)**	*(.98)*	*(.97)*	*(.99)*	*(.98)*			
11. English/Writing	**0.88**	**0.86**	**0.89**	**0.87**	**0.88**	*0.98*	*0.91*	*0.94*	*0.95*	*0.97*	*1.00*	
12. Writing	**0.68**	**0.67**	**0.71**	**0.70**	**0.68**	*0.77*	*0.74*	*0.73*	*0.78*	*0.77*	*0.88*	*1.00*

As can be seen in Table 5, correlations were all positive and very strong at the school level. There was little change from 2006 to 2007. The 2007 correlations may be slightly stronger than the 2006 results, but not so much so as to cause concern. This table shows very clearly that schools that performed well on one indicator of student performance tended to perform well on all of them for both 2006 and 2007, as did students (see Table 4).

Pooled within School Correlations

There is some concern with examining only overall correlations at the school level. It is possible that schools differ so much in terms of the overall student performance that correlations of components of PSAE might be large simply because of the large differences in overall performance from school to school. By examining the pooled-within-school correlations it is possible to examine whether the relations among components of PSAE changed from 2006 to 2007 within schools. These analyses guard against the possibility that the prior school-level correlations only capture the relative differences in schools overall. The pooled correlations are aggregate measures from each Illinois school. Table 6 contains the pooled correlations for 2006 and 2007. The formatting is the same as for prior correlation tables.

Table 6. 2007 Pooled Within School Correlations With Corresponding 2006 Values in Parentheses

Variable	1	2	3	4	5	6	7	8	9	10	11	12
WorkKeys												
1. Math Scale Score	*1.00*											
2. Math Level Score	0.93	*1.00*										
3. Reading Scale Score	*0.65*	*0.62*	*1.00*									
4. Reading Level Score	*0.60*	*0.59*	0.93	*1.00*								
		(.60)										
Internally Developed												
5. Science Scale Score	**0.71**	**0.66**	**0.65**	**0.60**	*1.00*							
		(.65)		**(.59)**								
ACT												
6. English	**0.69**	**0.64**	**0.68**	**0.62**	**0.70**	*1.00*						
		(.63)		**(.63)**	**(.68)**							
7. Math	**0.81**	**0.71**	**0.61**	**0.54**	**0.71**	*0.74*	*1.00*					
		(.68)		**(.54)**	**(.69)**	*(.73)*						
8. Reading	**0.62**	**0.56**	**0.66**	**0.59**	**0.67**	*0.78*	*0.65*	*1.00*				
		(.55)		**(.58)**	**(.66)**	*(.79)*	*(.67)*					
9. Science	**0.69**	**0.62**	**0.63**	**0.57**	**0.69**	*0.72*	*0.75*	*0.70*	*1.00*			
		(.63)		**(.58)**	**(.69)**	*(.74)*	*(.76)*	*(.72)*				
10. Composite	**0.78**	**0.71**	**0.72**	**0.65**	**0.77**	*0.92*	*0.87*	*0.89*	*0.88*	*1.00*		
		(.69)		**(.65)**	**(.76)**	*(.92)*	*(.87)*	*(.90)*	*(.89)*			
11. English/Writing	**0.69**	**0.64**	**0.69**	**0.64**	**0.68**	*0.96*	*0.72*	*0.76*	*0.71*	*0.89*	*1.00*	
12. Writing	**0.42**	**0.42**	**0.44**	**0.43**	**0.39**	*0.50*	*0.41*	*0.43*	*0.43*	*0.50*	*0.72*	*1.00*

If school-level differences on the PSAE components were substantial, then the pooled-within-school correlations would be higher than the corresponding correlations across all students (presented in Table 4). However, the pooled-within-school correlations for both the 2006 and 2007 administrations are consistently lower than the corresponding student-level correlations. The 2006 and 2007 pooled correlations are also similar in magnitude. Taken together, these analyses indicate that there were no school-level differences that were excessively contributing to the scoring patterns observed in the 2007 administration, but that rather these patterns are a statewide phenomenon.

To continue exploring any potential differences between years, correlations between the like components of the 2006 and 2007 administrations were computed. Table 7 presents these results formatted the same as prior correlation tables. In addition to the previously indicated formatting, Table 7 also contains highlighting to indicate the same test correlation across years (e.g. ACT Math 2006 with ACT Math 2007).

Table 7. School-Level Correlations Across Administration Years

	2007							
Variable	1	2	3	4	5	6	7	8
2006								
WorkKeys								
1. Math Level	*0.90*							
2. Reading Level	*0.79*	*0.77*						
Internally Developed								
3. Science Scale Score	**0.87**	**0.79**	*0.92*					
ACT								
4. English	**0.81**	**0.78**	**0.83**	*0.89*				
5. Math	**0.84**	**0.76**	**0.86**	*0.85*	*0.93*			
6. Reading	**0.83**	**0.79**	**0.87**	*0.87*	*0.87*	*0.89*		
7. Science	**0.84**	**0.78**	**0.88**	*0.87*	*0.88*	*0.88*	*0.88*	
8. Composite	**0.85**	**0.80**	**0.88**	*0.89*	*0.90*	*0.90*	*0.89*	*0.91*

The correlations between the 2006 and 2007 school-level scores on WorkKeys Reading from Table 7 are noticeably lower than the other correlations between like components across the administration years. The correlation is still positive and strong, but at 0.77, much lower than the next lowest component correlation across years, which is ACT Science at 0.88. This suggests that something unique to the 2007 WorkKeys Information Reading test constrained its correlation with the previous year's results.

Regression Analyses of Demographic Variables

Regression analysis is another way to look at across-year differences that might exist. In this case, school-level means from the 2006 administration were used to predict 2007 school-level means. Initial regression coefficients are the same as the previously calculated school-level correlations. The important piece of this regression analysis is that it allows us to determine whether student subgroups performed differentially better or worse on any PSAE component than would be expected, given the previous scoring pattern. To explore this possibility, initial regression equations were computed for 2007, based on 2006 score data. Then, school-level demographic characteristics were added to the equations to determine if taking into consideration a school's subgroup population proportions increased our ability to predict the 2007 score from 2006. If the predictability of the equation increased, that would suggest that the two tests were measuring student achievement differently, depending on a school's demographic makeup. Tables 8 through 15 present regression results for the three PSAE components.

Table 8. Regression Results Showing the Adjusted Strength of School-Level Demographic Characteristics on Predicting School-Level 2007 WorkKeys Reading Scores Based on School-Level 2006 WorkKeys Reading Scores

Gender			
Standardized Coefficient			
WorkKeys Reading 2006	Gender	R^2	Change in R^2 due to Demographic
0.774		0.599	
0.779	0.038	0.599	0.000
Ethnicity (African American)			
Standardized Coefficient			
WorkKeys Reading 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.774		0.599	
0.694	-0.142	0.611	0.012
Ethnicity (Hispanic)			
Standardized Coefficient			
WorkKeys Reading 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.774		0.599	
0.737	-0.075	0.584	-0.015
Economic Disadvantage			
Standardized Coefficient			
WorkKeys Reading 2006	Economically Disadvantaged	R^2	Change in R^2 due to Demographic
0.774		0.599	
0.583	-0.295	0.649	0.050
Students with Disabilities			
Standardized Coefficient			
WorkKeys Reading 2006	Students with Disabilities	R^2	Change in R^2 due to Demographic
0.774		0.599	
0.719	-0.213	0.640	0.041
Limited English Proficiency			
Standardized Coefficient			
WorkKeys Reading 2006	Limited English Proficiency	R^2	Change in R^2 due to Demographic
0.774		0.599	
0.771	-0.028	0.599	0.000

Table 9. Regression Results Showing the Adjusted Strength of School-Level Demographic Characteristics on Predicting School-Level 2007 WorkKeys Math Scores Based on School-Level 2006 WorkKeys Math Scores

Gender			
Standardized Coefficient			
WorkKeys Math 2006	Gender	R^2	Change in R^2 due to Demographic
0.902		0.813	
0.899	-0.013	0.813	0.000
Ethnicity (African American)			
Standardized Coefficient			
WorkKeys Math 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.902		0.813	
0.802	-0.137	0.822	0.009
Ethnicity (Hispanic)			
Standardized Coefficient			
WorkKeys Math 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.902		0.813	
0.888	-0.014	0.796	-0.017
Economic Disadvantage			
Standardized Coefficient			
WorkKeys Math 2006	Economically Disadvantaged	R^2	Change in R^2 due to Demographic
0.902		0.813	
0.757	-0.191	0.828	0.015
Students with Disabilities			
Standardized Coefficient			
WorkKeys Math 2006	Students with Disabilities	R^2	Change in R^2 due to Demographic
0.902		0.813	
0.865	-0.130	0.828	0.015
Limited English Proficiency			
Standardized Coefficient			
WorkKeys Math 2006	Limited English Proficiency	R^2	Change in R^2 due to Demographic
0.902		0.813	
0.900	-0.016	0.813	0.000

Table 10. Regression Results Showing the Adjusted Strength of School-Level Demographic Characteristics on Predicting School-Level 2007 Science Scale Scores Based on School-Level 2006 Science Scale Scores

Gender			
Standardized Coefficient			
Science Scale Score 2006	Gender	R^2	Change in R^2 due to Demographic
0.918		0.842	
0.914	-0.018	0.842	0.000
Ethnicity (African American)			
Standardized Coefficient			
Science Scale Score 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.918		0.842	
0.856	-0.084	0.844	0.002
Ethnicity (Hispanic)			
Standardized Coefficient			
Science Scale Score 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.918		0.842	
0.900	-0.029	0.829	-0.013
Economic Disadvantage			
Standardized Coefficient			
Science Scale Score 2006	Economically Disadvantaged	R^2	Change in R^2 due to Demographic
0.918		0.842	
0.750	-0.212	0.858	0.016
Students with Disabilities			
Standardized Coefficient			
Science Scale Score 2006	Students with Disabilities	R^2	Change in R^2 due to Demographic
0.918		0.842	
0.887	-0.116	0.854	0.012
Limited English Proficiency			
Standardized Coefficient			
Science Scale Score 2006	Limited English Proficiency	R^2	Change in R^2 due to Demographic
0.918		0.842	
0.915	-0.018	0.842	0.000

Table 11. Regression Results Showing the Adjusted Strength of School-Level Demographic Characteristics on Predicting School-Level 2007 ACT English Scores Based on School-Level 2006 ACT English Scores

Gender			
Standardized Coefficient			
ACT English 2006	Gender	R^2	Change in R^2 due to Demographic
0.885		0.782	
0.890	0.047	0.784	0.002
Ethnicity (African American)			
Standardized Coefficient			
ACT English 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.885		0.782	
0.834	-0.095	0.788	0.006
Ethnicity (Hispanic)			
Standardized Coefficient			
ACT English 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.885		0.782	
0.875	-0.020	0.775	-0.007
Economic Disadvantage			
Standardized Coefficient			
ACT English 2006	Economically Disadvantaged	R^2	Change in R^2 due to Demographic
0.885		0.782	
0.769	-0.175	0.799	0.017
Students with Disabilities			
Standardized Coefficient			
ACT English 2006	Students with Disabilities	R^2	Change in R^2 due to Demographic
0.885		0.782	
0.852	-0.123	0.769	-0.013
Limited English Proficiency			
Standardized Coefficient			
ACT English 2006	Limited English Proficiency	R^2	Change in R^2 due to Demographic
0.885		0.782	
0.884	-0.015	0.782	0.000

Table 12. Regression Results Showing the Adjusted Strength of School-Level Demographic Characteristics on Predicting School-Level 2007 ACT Math Scores Based on School-Level 2006 ACT Math Scores

Gender			
Standardized Coefficient			
ACT Math 2006	Gender	R^2	Change in R^2 due to Demographic
0.925		0.855	
0.925	0.005	0.855	0.000
Ethnicity (African American)			
Standardized Coefficient			
ACT Math 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.925		0.855	
0.892	-0.057	0.857	0.002
Ethnicity (Hispanic)			
Standardized Coefficient			
ACT Math 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.925		0.855	
0.918	-0.010	0.846	-0.009
Economic Disadvantage			
Standardized Coefficient			
ACT Math 2006	Economically Disadvantaged	R^2	Change in R^2 due to Demographic
0.925		0.855	
0.840	-0.120	0.862	0.007
Students with Disabilities			
Standardized Coefficient			
ACT Math 2006	Students with Disabilities	R^2	Change in R^2 due to Demographic
0.925		0.855	
0.901	-0.087	0.862	0.007
Limited English Proficiency			
Standardized Coefficient			
ACT Math 2006	Limited English Proficiency	R^2	Change in R^2 due to Demographic
0.925		0.855	
0.923	-0.015	0.855	0.000

Table 13. Regression Results Showing the Adjusted Strength of School-Level Demographic Characteristics on Predicting School-Level 2007 ACT Reading Scores Based on School-Level 2006 ACT Reading Scores

Gender			
Standardized Coefficient			
ACT Reading 2006	Gender	R^2	Change in R^2 due to Demographic
0.888		0.788	
0.890	0.017	0.788	0.000
Ethnicity (African American)			
Standardized Coefficient			
ACT Reading 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.888		0.788	
0.805	-0.143	0.801	0.013
Ethnicity (Hispanic)			
Standardized Coefficient			
ACT Reading 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.888		0.788	
0.865	-0.056	0.780	-0.008
Economic Disadvantage			
Standardized Coefficient			
ACT Reading 2006	Economically Disadvantaged	R^2	Change in R^2 due to Demographic
0.888		0.788	
0.725	-0.234	0.816	0.028
Students with Disabilities			
Standardized Coefficient			
ACT Reading 2006	Students with Disabilities	R^2	Change in R^2 due to Demographic
0.888		0.788	
0.859	-0.117	0.801	0.013
Limited English Proficiency			
Standardized Coefficient			
ACT Reading 2006	Limited English Proficiency	R^2	Change in R^2 due to Demographic
0.888		0.788	
0.886	-0.020	0.788	0.000

Table 14. Regression Results Showing the Adjusted Strength of School-Level Demographic Characteristics on Predicting School-Level 2007 ACT Science Scores Based on School-Level 2006 ACT Science Scores

Gender			
Standardized Coefficient			
ACT Science 2006	Gender	R^2	Change in R^2 due to Demographic
0.883		0.779	
0.884	0.011	0.778	-0.001
Ethnicity (African American)			
Standardized Coefficient			
ACT Science 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.883		0.779	
0.836	-0.078	0.782	0.003
Ethnicity (Hispanic)			
Standardized Coefficient			
ACT Science 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.883		0.779	
0.877	-0.009	0.772	-0.007
Economic Disadvantage			
Standardized Coefficient			
ACT Science 2006	Economically Disadvantaged	R^2	Change in R^2 due to Demographic
0.883		0.779	
0.754	-0.181	0.795	0.016
Students with Disabilities			
Standardized Coefficient			
ACT Science 2006	Students with Disabilities	R^2	Change in R^2 due to Demographic
0.883		0.779	
0.847	-0.138	0.796	0.017
Limited English Proficiency			
Standardized Coefficient			
ACT Science 2006	Limited English Proficiency	R^2	Change in R^2 due to Demographic
0.883		0.779	
0.881	-0.017	0.779	0.000

Table 15. Regression Results Showing the Adjusted Strength of School-Level Demographic Characteristics on Predicting School-Level 2007 ACT Composite Scores Based on School-Level 2006 ACT Composite Scores

Gender			
Standardized Coefficient			
ACT Composite 2006	Gender	R^2	Change in R^2 due to Demographic
0.912		0.832	
0.915	0.022	0.832	0.000
Ethnicity (African American)			
Standardized Coefficient			
ACT Composite 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.912		0.832	
0.863	-0.085	0.836	0.004
Ethnicity (Hispanic)			
Standardized Coefficient			
ACT Composite 2006	Ethnicity	R^2	Change in R^2 due to Demographic
0.912		0.832	
0.903	-0.019	0.825	-0.007
Economic Disadvantage			
Standardized Coefficient			
ACT Composite 2006	Economically Disadvantaged	R^2	Change in R^2 due to Demographic
0.912		0.832	
0.795	-0.166	0.846	0.014
Students with Disabilities			
Standardized Coefficient			
ACT Composite 2006	Students with Disabilities	R^2	Change in R^2 due to Demographic
0.912		0.832	
0.882	-0.113	0.844	0.012
Limited English Proficiency			
Standardized Coefficient			
ACT Composite 2006	Limited English Proficiency	R^2	Change in R^2 due to Demographic
0.912		0.832	
0.911	-0.016	0.832	0.000

Based on Tables 8-15, for most PSAE subtests, adding school-level demographic characteristics to the regression equation predicting 2007 results from 2006 added less than 2% to the overall prediction (based on R^2). This does not mean that there were not substantial performance gaps among subgroups, only that using prior performance as a predictor, in most instances, accounted for the lion's share of the variability. There were several PSAE component tests for which adding demographic information accounted for between 1% and 2 % of the overall variability in scores. These results are not particularly concerning by themselves but will be considered in conjunction with effect size statistics later in the report.

School-level subgroup proportions, however, did account for 4% to 5% of the overall variability in scores not accounted for by previous performance in two of the regression equations. Adding the proportions of students with disabilities improved the school-level prediction for WorkKeys Reading by 4.1%. Adding the proportion of economically disadvantaged students added 5%. The direction of these regression equation results suggests a differential impact for WorkKeys Reading in 2007 (test scores may have been differentially lower for students in these subgroups in 2007 compared to 2006). Again, this result in isolation does not necessarily indicate test bias, but it should be considered as an indicator of potential test bias and weighed in combination with effect size analyses.

Effect Size Analyses of Demographic Variables

Since the results for WorkKeys Reading declines from 2006 to 2007 while the other student performance measures increase, it is important to investigate the impact of that difference on student subgroups. Regression analyses at the school level indicated a possible connection between scoring patterns on WorkKeys Reading and the proportions of students identified as economically disadvantaged or those with disabilities. Effect size statistics can help us ascertain if membership in a particular subgroup has impacted the scoring gaps among the various student subgroups. Tables 16-21 contain effect size statistics in a common and comparable metric for 2006 and 2007. When examining these tables, it is important to keep in mind that while many achievement gaps are large, the purpose of these analyses is to compare one year to the other.

Table 16. Effect Size Statistics for Gender on PSAE Components

Year	Component	Gender	Mean	Standard Deviation	Number of Cases	Effect Size *d*
2007	WorkKeys Math	Male	4.74	1.803	64,700	
	WorkKeys Math	Female	4.53	1.732	67,187	0.06
	WorkKeys Reading	Male	4.35	1.527	64,677	
	WorkKeys Reading	Female	4.58	1.259	67,177	-0.08
	Science Scale Score	Male	71.56	10.176	64,732	
	Science Scale Score	Female	69.33	8.892	67,215	0.12
	ACT English	Male	18.87	6.354	64,706	
	ACT English	Female	20.15	6.340	67,212	-0.10
	ACT Math	Male	20.57	5.743	64,697	
	ACT Math	Female	19.80	5.219	67,206	0.07
	ACT Reading	Male	19.47	6.152	64,657	
	ACT Reading	Female	20.18	6.058	67,187	-0.06
	ACT Science	Male	20.02	5.465	64,634	
	ACT Science	Female	19.72	4.723	67,175	0.03
2006	WorkKeys Math	Male	4.70	1.768	61,218	
	WorkKeys Math	Female	4.53	1.687	63,832	0.05
	WorkKeys Reading	Male	4.57	1.614	61,191	
	WorkKeys Reading	Female	4.83	1.314	63,826	-0.09
	Science Scale Score	Male	72.01	10.081	61,233	
	Science Scale Score	Female	69.16	8.637	63,836	0.15
	ACT English	Male	18.93	6.322	61,011	
	ACT English	Female	19.95	6.129	63,530	-0.08
	ACT Math	Male	20.28	5.596	61,010	
	ACT Math	Female	19.57	5.058	63,527	0.07
	ACT Reading	Male	19.67	6.437	60,986	
	ACT Reading	Female	20.02	6.196	63,520	-0.03
	ACT Science	Male	20.00	5.386	60,964	
	ACT Science	Female	19.53	4.879	63,515	0.05

Table 17. Effect Size Statistics for Ethnicity (African American) on PSAE Components

Year	Component	Ethnicity	Mean	Standard Deviation	Number of Cases	Effect Size *d*
2007	WorkKeys Math	White	5.02	1.584	87,032	
	WorkKeys Math	African American	3.32	1.819	21,030	0.45
	WorkKeys Reading	White	4.67	1.328	87,018	
	WorkKeys Reading	African American	3.88	1.451	21,018	0.27
	Science Scale Score	White	72.76	9.180	87,049	
	Science Scale Score	African American	63.47	7.585	21,041	0.48
	ACT English	White	20.87	6.251	87,093	
	ACT English	African American	15.67	5.052	20,959	0.42
	ACT Math	White	21.30	5.520	87,084	
	ACT Math	African American	16.51	3.366	20,955	0.46
	ACT Reading	White	21.23	6.070	87,056	
	ACT Reading	African American	15.99	4.447	20,938	0.44
	ACT Science	White	20.90	5.062	87,036	
	ACT Science	African American	16.73	3.915	20,930	0.42
2006	WorkKeys Math	White	4.97	1.539	81,556	
	WorkKeys Math	African American	3.39	1.835	18,907	0.42
	WorkKeys Reading	White	4.91	1.395	81,536	
	WorkKeys Reading	African American	4.09	1.534	18,896	0.27
	Science Scale Score	White	72.86	9.041	81,564	
	Science Scale Score	African American	63.42	7.448	18,910	0.50
	ACT English	White	20.67	6.15	81,446	
	ACT English	African American	15.84	4.97	18,549	0.40
	ACT Math	White	20.97	5.343	81,442	
	ACT Math	African American	16.32	3.287	18,550	0.46
	ACT Reading	White	21.1	6.291	81,433	
	ACT Reading	African American	16.13	4.705	18,539	0.41
	ACT Science	White	20.8	5.054	81,419	
	ACT Science	African American	16.53	3.947	18,531	0.43

Table 18. Effect Size Statistics for Ethnicity (Hispanic) on PSAE Components

Year	Component	Ethnicity	Mean	Standard Deviation	Number of Cases	Effect Size *d*
2007	WorkKeys Math	White	5.02	1.584	87,032	
	WorkKeys Math	Hispanic	4.00	1.703	16,784	0.30
	WorkKeys Reading	White	4.67	1.328	87,018	
	WorkKeys Reading	Hispanic	3.99	1.432	16,777	0.24
	Science Scale Score	White	72.76	9.180	87,049	
	Science Scale Score	Hispanic	65.84	8.003	16,815	0.37
	ACT English	White	20.87	6.251	87,093	
	ACT English	Hispanic	16.42	5.211	16,783	0.36
	ACT Math	White	21.30	5.520	87,084	
	ACT Math	Hispanic	17.73	3.985	16,781	0.35
	ACT Reading	White	21.23	6.070	87,056	
	ACT Reading	Hispanic	16.79	4.785	16,769	0.38
	ACT Science	White	20.90	5.062	87,036	
	ACT Science	Hispanic	17.61	4.087	16,765	0.34
2006	WorkKeys Math	White	4.97	1.539	81,556	
	WorkKeys Math	Hispanic	3.99	1.719	14,823	0.29
	WorkKeys Reading	White	4.91	1.395	81,536	
	WorkKeys Reading	Hispanic	4.20	1.501	14,824	0.24
	Science Scale Score	White	72.86	9.041	81,564	
	Science Scale Score	Hispanic	6.07	7.999	14,830	0.37
	ACT English	White	20.67	6.15	81,446	
	ACT English	Hispanic	16.35	5.114	14,782	0.36
	ACT Math	White	20.97	5.343	81,442	
	ACT Math	Hispanic	17.45	3.798	14,782	0.35
	ACT Reading	White	21.1	6.291	81,433	
	ACT Reading	Hispanic	16.90	5.077	14,776	0.34
	ACT Science	White	20.8	5.054	81,419	
	ACT Science	Hispanic	17.41	4.174	14,772	0.34

Table 19. Effect Size Statistics for Economic Disadvantage Status on PSAE Components

Year	Component	Economic Disadvantage Status	Mean	Standard Deviation	Number of Cases	Effect Size *d*
2007	WorkKeys Math	Not Eligible	4.98	1.610	95,794	
	WorkKeys Math	Eligible	3.70	1.838	36,085	0.35
	WorkKeys Reading	Not Eligible	4.67	1.311	95,783	
	WorkKeys Reading	Eligible	3.93	1.489	36,063	0.26
	Science Scale Score	Not Eligible	72.52	9.231	95,817	
	Science Scale Score	Eligible	64.86	8.275	36,122	0.40
	ACT English	Not Eligible	20.86	6.271	95,857	
	ACT English	Eligible	15.97	5.192	36,021	0.39
	ACT Math	Not Eligible	21.32	5.591	95,848	
	ACT Math	Eligible	17.15	3.837	36,015	0.40
	ACT Reading	Not Eligible	21.08	6.090	95,821	
	ACT Reading	Eligible	16.51	4.791	35,983	0.38
	ACT Science	Not Eligible	20.87	5.063	95,797	
	ACT Science	Eligible	17.20	4.160	35,972	0.37
2006	WorkKeys Math	Not Eligible	4.94	1.561	89,522	
	WorkKeys Math	Eligible	3.7	1.827	32,141	0.34
	WorkKeys Reading	Not Eligible	4.92	1.382	89,504	
	WorkKeys Reading	Eligible	4.11	1.553	32,129	0.27
	Science Scale Score	Not Eligible	72.64	9.106	89,533	
	Science Scale Score	Eligible	64.87	8.085	32,148	0.41
	ACT English	Not Eligible	20.72	6.124	89,385	
	ACT English	Eligible	15.93	5.099	31,767	0.39
	ACT Math	Not Eligible	21.02	5.414	89,383	
	ACT Math	Eligible	16.86	3.673	31,766	0.41
	ACT Reading	Not Eligible	21.09	6.293	89,366	
	ACT Reading	Eligible	16.41	4.946	31,755	0.38
	ACT Science	Not Eligible	20.79	5.06	89,350	
	ACT Science	Eligible	16.93	4.196	31,744	0.38

Table 20. Effect Size Statistics for Disability Status on PSAE Components

Year	Component	Disability Status	Mean	Standard Deviation	Number of Cases	Effect Size *d*
2007	WorkKeys Math	No Disability	4.87	1.573	116,743	
	WorkKeys Math	Disability	2.74	2.052	15,136	0.50
	WorkKeys Reading	No Disability	4.66	1.186	116,732	
	WorkKeys Reading	Disability	2.98	1.944	15,114	0.46
	Science Scale Score	No Disability	71.63	9.003	116,756	
	Science Scale Score	Disability	61.16	9.046	15,183	0.50
	ACT English	No Disability	20.34	6.077	116,659	
	ACT English	Disability	13.30	5.066	15,219	0.53
	ACT Math	No Disability	20.77	5.425	116,653	
	ACT Math	Disability	15.65	3.592	15,210	0.49
	ACT Reading	No Disability	20.41	5.981	116,615	
	ACT Reading	Disability	15.42	5.272	15,189	0.40
	ACT Science	No Disability	20.45	4.904	116,604	
	ACT Science	Disability	15.35	4.271	15,165	0.48
2006	WorkKeys Math	No Disability	4.85	1.53	108,275	
	WorkKeys Math	Disability	2.71	2.013	13,388	0.51
	WorkKeys Reading	No Disability	4.92	1.261	108,262	
	WorkKeys Reading	Disability	3.01	1.902	13,371	0.51
	Science Scale Score	No Disability	71.72	8.952	108,279	
	Science Scale Score	Disability	61.49	8.758	13,402	0.50
	ACT English	No Disability	20.25	5.93	107,716	
	ACT English	Disability	13.11	4.865	13,436	0.55
	ACT Math	No Disability	20.48	5.281	107,716	
	ACT Math	Disability	15.48	3.349	13,433	0.49
	ACT Reading	No Disability	20.43	6.229	107,705	
	ACT Reading	Disability	15.29	4.989	13,416	0.41
	ACT Science	No Disability	20.34	4.97	107,695	
	ACT Science	Disability	15.28	4.148	13,399	0.48

Table 21. Effect Size Statistics for English Language Status on PSAE Components

Year	Component	English Language Status	Mean	Standard Deviation	Number of Cases	Effect Size *d*
2007	WorkKeys Math	Non ELL	4.64	1.765	130,285	
	WorkKeys Math	ELL	3.77	1.948	1,594	0.23
	WorkKeys Reading	Non ELL	4.48	1.393	130,254	
	WorkKeys Reading	ELL	3.58	1.727	1,592	0.28
	Science Scale Score	Non ELL	70.50	9.590	130,342	
	Science Scale Score	ELL	64.54	9.249	1,597	0.30
	ACT English	Non ELL	19.57	6.367	130,275	
	ACT English	ELL	15.73	6.173	1,603	0.29
	ACT Math	Non ELL	20.21	5.499	130,260	
	ACT Math	ELL	17.93	4.704	1,603	0.22
	ACT Reading	Non ELL	19.87	6.111	130,202	
	ACT Reading	ELL	16.41	5.429	1,602	0.29
	ACT Science	Non ELL	19.89	5.101	130,167	
	ACT Science	ELL	17.62	4.628	1,602	0.23
2006	WorkKeys Math	Non ELL	4.62	1.723	120,907	
	WorkKeys Math	ELL	3.82	1.985	756	0.21
	WorkKeys Reading	Non ELL	4.71	1.468	120,878	
	WorkKeys Reading	ELL	3.63	1.859	755	0.31
	Science Scale Score	Non ELL	70.62	9.478	120,925	
	Science Scale Score	ELL	65.27	9.447	756	0.27
	ACT English	Non ELL	19.49	6.233	120,399	
	ACT English	ELL	15.34	5.687	753	0.33
	ACT Math	Non ELL	19.94	5.342	120,396	
	ACT Math	ELL	17.88	4.568	753	0.20
	ACT Reading	Non ELL	19.89	6.312	120,368	
	ACT Reading	ELL	15.98	5.371	753	0.32
	ACT Science	Non ELL	19.79	5.134	120,342	
	ACT Science	ELL	16.91	4.758	752	0.28

Tables 16 through 21 show sizeable gaps among subgroups identified for both 2006 and 2007 on all PSAE components. The gaps are relatively stable across measures and are very stable across years. This indicates that the decline in WorkKeys Reading scores does not seem to have unduly impacted any particular subgroup of students. The effect size statistics for all subgroups calculated for 2006 are nearly identical to those calculated for 2007. In the few instances where there are discernable differences, these small differences exhibit no consistent pattern. This also means that gaps among student subgroups remained largely unchanged between 2006 and 2007.

Conclusions and Discussion

WorkKeys Reading declined dramatically while other PSAE components improved from 2006 to 2007. This study verified and quantified the magnitude of that decline, but while several potential reasons for the decline were investigated, all were ultimately discarded. The decline was exhibited throughout the WorkKeys scale. It does not appear to have altered relations among the components of PSAE. Correlations were stable from year to year. The changes seem to have impacted all (or nearly all) schools in the same manner. Correlations of change scores remained stable across all PSAE components including WorkKeys Reading. No particular subgroup of students appears to have been differentially impacted by the decline. Gaps among the subgroups identified by NCLB remained very stable from 2006 to 2007 for all PSAE components.

The decline in WorkKeys reading was large (0.15 standard deviations). It is inconsistent with the concordance tables published by ACT linking WorkKeys Reading with ACT Reading. The ACT Reading scale scores most closely associated with WorkKeys reading levels increased by 1 to 2 points for all reporting levels from 2006 to 2007. The most obvious next step is to determine whether this decline was a national phenomenon or isolated to Illinois. If isolated to Illinois, investigations should be conducted to determine why the discrepancy might exist. It is possible that something changed for the Illinois WorkKeys Reading test in 2007 to cause the decline. Changing forms, shifting content (or tested standards), equating issues (e.g. processing errors, item position effects, or contextual effects), resetting cut scores, and other possible changes might help account for the decline. At this stage, however, we are unaware that any of these changes took place in 2007. If the pattern exists at the national level, it will be necessary to recreate the concordance tables. A deeper investigation of why the decline occurred may also be necessary.

It is certainly possible for scores on one kind of test to decline while others improve as curriculum and instructional priorities shift. These changes, however, typically take several years to manifest to the extent we see here. It is a slow and difficult process to implement large-scale changes in curriculum and instruction. NCLB has pushed states to work ever more diligently toward improving test scores, yet for most states, scores increase only a small amount each year and follow very predictable patterns, despite great efforts to alter and improve student achievement. This decline may be a one-year anomaly and scores may correct themselves in 2008. Test scores are estimates of student achievement or ability and those estimates certainly have an error component. They can bounce around a little. However, because the stakes for the test are so high under NCLB, the ramifications for this anomaly are great for districts, schools, education staff within the schools, and even students. All reasonable efforts should be taken to ensure that the scoring pattern exhibited for WorkKeys Reading truly represents a decline in student achievement rather than a statistical artifact or processing error.

One potential explanation for the PSAE scoring pattern observed between 2006 and 2007 relates to the equating of earlier administrations of the WorkKeys assessment. In 2004 and 2005, WorkKeys reading was equated using a random equivalent groups design. We do not know a great deal more about the design, however, including how the groups were selected or created. Establishing equivalent groups is vital for ensuring equating stability using this method. WorkKeys Reading means for Illinois increased by a substantial margin from 2004 to 2005. In fact the scores increased by almost exactly the same amount that they decreased from 2006 to

2007 (based on estimating effect size from the mean shift from 2004 to 2005 and current variance data). When asked about this increase and the equating methodology, ACT informed HumRRO that the 2007 WorkKeys Reading assessment was equated to the 2004 administration, while the 2006 assessment was equated to the 2005 administration using a common-item non-equivalent groups design (where some items are repeated from an earlier administration and equating is accomplished using the change in scoring patterns from one administration to the next). Since the 2004 and 2005 assessments were subjected to an equating procedure, it should not have mattered which earlier administration was used for equating 2006 and 2007. However, without further information about the 2004/2005 equating procedure we can not know how well that procedure worked. It is possible that some methodological or random error might have contributed significantly to the apparent gain from 2004 to 2005. Then, when the 2006 administration is equated to 2005 and the 2007 administration to 2004 using common item equating, that same methodological or random equating error would result in an apparent decrease in mean scores. HumRRO has no means of investigating the likelihood of this possibility from the data provided, but the overall data patterns indicate that this is a plausible explanation for the decline.

References

Bacci E. D., Koger, M. E., Hoffman, R. G., & Thacker, A. A. (2003). *Relationships among Kentucky's core content test, ACT scores, and students' self-reported high school grades for the classes of 2000 through 2002.* (HumRRO Draft Report FR-03-19). Louisville, KY: Human Resources Research Organization.

Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools. How all fifty states are above the national average.* Daniels, WV: Friends for Education.

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests.* Albuquerque, NM: Friends for Education.

Center on Education Policy (2007). *Answering the question that matters most: has student achievement increased since No Child Left Behind?* Available at www.cep-dc.org: Washington, D.C.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Hoffman, R. G., & Tannen, M. B. (1998). *Relationships between Kentucky's open-response scores for eighth grade students and their CTBS-5 scores as ninth grade students* (HumRRO Report No. FR-WATSD-98-30). Radcliff, KY: Human Resources Research Organization.

Linn, R. L. (1998). *Assessments and accountability*. Center for the Study of Evaluation (CSE) Report 490. National Center for Research on Evaluation, Standards and Student Testing (CRESST): Los Angeles, CA.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1435 (2002).