NATIONAL CAREER READINESS CERTIFICATE™

WorkKeys® Assessments Technical Bulletin

ACT®

Contents

Introduction 1
National Career Readiness Certificate 3
National Career Readiness Certificate Skill Levels 3
Foundations for the Content of the WorkKeys System 4
Reliability 5
Internal Consistency for Number-Correct (NC) Scores 5
Generalizability Analyses 6
Standard Error of Measurement for Number-Correct Scores and Scale Scores 8
Classification Consistency for Level Scores 12
Scaling and Equating 16
Level Score Scale 16
Scale Scores 29
Equating 31
WorkKeys Validity Evidence 34
Construct-Related Evidence 35
Criterion-Related Evidence 41
Content-Related Evidence 46
Adverse Impact 48
WorkKeys Job Analysis Options 52
References 54

List of Tables

Table 1 Estimated Variance Components, Error Variances, and Generalizability Coefficients—*Reading for Information* 7
Table 2 Estimated Variance Components, Error Variances, and Generalizability Coefficients—*Applied Mathematics* 7
Table 3 Estimated Variance Components, Error Variances, and Generalizability Coefficients—*Locating Information* 8
Table 4 Predicted Classification Consistency for Level Scores—*Reading for Information* 13
Table 5 Predicted Classification Consistency for Level Scores—*Applied Mathematics* 14
Table 6 Predicted Classification Consistency for Level Scores—*Locating Information* 15
Table 7 Summary Statistics for *Reading for Information* NC Scores 17
Table 8 Boundary Thetas, Form Cutoff Thetas, and NC Score Cutoffs—*Reading for Information* 19
Table 9 Percentage of Test Takers by Level Scores by Form—*Reading for Information* 21
Table 10 Summary Statistics for *Applied Mathematics* NC Scores 22
Table 11 Boundary Thetas, Form Cutoff Thetas, and NC Score Cutoffs—*Applied Mathematics* 24
Table 12 Percentage of Test Takers by Level Scores by Form—*Applied Mathematics* 25
Table 13 Summary Statistics for *Locating Information* NC Scores 26
Table 14 Boundary Thetas, Form Cutoff Thetas, and NC Score Cutoffs—*Locating Information* 27
Table 15 Percentage of Test Takers by Level Scores by Form—*Locating Information* 29
Table 16 Summary Statistics for *Reading for Information* NC Scores and Scale Scores 30

Table 17 Summary Statistics for *Applied Mathematics* NC Scores and Scale Scores 31
Table 18 Summary Statistics for *Locating Information* NC Scores and Scale Scores 31
Table 19 Correlations between WorkKeys *Reading for Information*, ACT Reading, and ACT English 35
Table 20 Percents of Test Takers by WorkKeys *Reading for Information* Level Scores and Ranges of ACT Reading Scale Scores 36
Table 21 WorkKeys *Applied Mathematics* and ACT Mathematics Score Correlations 38
Table 22 Percents of Test Takers by WorkKeys *Applied Mathematics* Level Scores and Ranges of ACT Mathematics Scale Scores . . 39
Table 23 Correlations between WorkKeys *Reading for Information* Scores and Job Performance Ratings 42
Table 24 Correlations between WorkKeys *Applied Mathematics* Level Scores and Job Performance Ratings 43
Table 25 Correlations between WorkKeys *Locating Information* Level Scores and Job Performance Ratings 43
Table 26 Job Classification Consistency with *Reading for Information* . . . 44
Table 27 Job Classification Consistency with *Applied Mathematics* 45
Table 28 Job Classification Consistency with *Locating Information* 45
Table 29 Descriptive Statistics of *Reading for Information* Mean Level Scores by Gender and Race/Ethnicity 49
Table 30 Descriptive Statistics of *Applied Mathematics* Mean Level Scores by Gender and Race/Ethnicity 50
Table 31 Descriptive Statistics of *Locating Information* Mean Level Scores by Gender and Race/Ethnicity 51

List of Figures

Figure 1 WorkKeys Assessments: General Description 2
Figure 2 SEMs for Two Forms of *Reading for Information* 9
Figure 3 SEMs for Two Forms of *Applied Mathematics* 10
Figure 4 SEMs for Three Forms of *Locating Information* 11
Figure 5 Item p-Values (p) and Mean Item p-Values (Connected) by Level of Item—*Reading for Information* 17
Figure 6 *Reading for Information* Level Characteristic Curves 18
Figure 7 Item p-Values (p) and Mean Item p-Values (Connected) by Level of Item—*Applied Mathematics* 23
Figure 8 *Applied Mathematics* Level Characteristic Curves 23
Figure 9 Item p-Values (p) and Mean Item p-Values (Connected) by Level of Item—*Locating Information* 26
Figure 10 *Locating Information* Level Characteristic Curves 27
Figure 11 Boxplots of Scale Scores on ACT Reading at Each Level Score on WorkKeys *Reading for Information* 37
Figure 12 Boxplots of Scale Scores on ACT Mathematics at Each Level Score on WorkKeys *Applied Mathematics* 40
Figure 13 Comparison of the *Uniform Guidelines* Requirements and Two ACT WorkKeys Job Analysis Procedures for Content Validation 47

Introduction

Economic development in local, state, and regional areas, once driven by demands closer to home, is now affected by a global economy. Further, faced with the fast-paced evolution of technology, employers are looking for employees who have the necessary skills to perform the jobs of today and to adapt to the jobs of tomorrow. Today's workforce must be able to increase their skills to sustain quality performance as required by rapidly changing jobs in a changing economy.

Individuals moving into the workforce, changing jobs or careers, or returning to the job market after an extended absence need to show potential employers evidence of their employability. Over 60% of people leaving high school are in the labor market within a year after leaving or graduating from school (U.S. Department of Education National Center for Education Statistics, 2003; U.S. Department of Labor—Bureau of Labor Statistics, 2003). "The average person born in the later years of the baby boom [1957–1964] held nearly 10 jobs from ages 18 to 36. More than two-thirds of these jobs were held in the first half of the period, from ages 18 to 27" (U.S. Department of Labor—Bureau of Labor Statistics, 2002). With a civilian labor force of about 145 million, an estimated 5.5 to 6 million people (about 4%) are likely to be changing jobs at any given time. Most of these people have relatively short records of performance, yet they need to communicate their employability characteristics to their potential employers. An unbiased assessment of essential skills yields important information for these candidates to present to employers for review or to use in improving their job-readiness skills.

> The vast majority of American manufacturers are experiencing a serious shortage of qualified employees, which in turn is causing significant impact to business and the ability of the country as a whole to compete in a global economy. This is the key finding of the 2005 Skills Gap Survey (Deloitte Development, 2005).

WorkKeys® is a foundational skills assessment system for measuring real-world skills critical to job success. WorkKeys can be used by employers, educators, and training organizations to meet skills assessment needs. At the same time, WorkKeys provides a common language for use by all stakeholders. The system includes assessments, job analysis, and training/instructional support, which together enable users to identify skill gaps and training needs and respond accordingly.

As a part of the WorkKeys system, ACT has profiled more than 13,000 individual jobs across the country to determine the skills and skill levels needed to succeed in them. According to our findings, three skills are highly important to most jobs.

- **Reading for Information**
- **Applied Mathematics**
- **Locating Information**

The WorkKeys tests measuring these skills are described in Figure 1.

Figure 1
WorkKeys Assessments: General Description

Reading for Information

The WorkKeys *Reading for Information* test measures the skills people use when they read and use written text in order to do a job. The written texts include memos, letters, directions, signs, notices, bulletins, policies, and regulations, based on materials that reflect actual reading demands of the workplace.

Applied Mathematics

The WorkKeys *Applied Mathematics* test measures the skills people use when they apply mathematical reasoning and problem-solving techniques to work-related problems. The test questions require the test taker to set up and solve the types of problems and do the types of calculations that actually occur in the workplace.

Locating Information

The WorkKeys *Locating Information* test measures the locating, comparative, summarization, and analytic skills people use when they work with workplace graphics such as charts, graphs, tables, forms, flowcharts, diagrams, floor plans, maps, and instrument gauges.

National Career Readiness Certificate

A solid foundation in these three skills is essential for a well-qualified workforce. Thus, these skills form the basis for ACT's National Career Readiness Certificate.

The National Career Readiness System links qualified individuals with employers who recognize the value of skilled job applicants. This comprehensive employment tool—available via the Internet—offers four components.

1. **Certification:** The National Career Readiness Certificate verifies at what level an individual has the foundational skills that are essential to performance on the job or in a training program.
2. **Certificate Registry:** This Internet-based system allows an individual to view WorkKeys scores, apply for a certificate, and order paper copies, as well as enabling employers to verify that an individual has a certificate.
3. **Talent Bank:** Individuals who qualify for a National Career Readiness Certificate can use the Talent Bank to post credentials for employers and search job postings in a national job database.
4. **Job Bank:** Employers who accept the National Career Readiness Certificate can post job opportunities and search for qualified candidates.

Because the certificate validates that an individual has certain essential skills important across a range of jobs, employers, job seekers, economic development professionals, and educators can use the certificate as a common language to improve the quality of the workforce.

National Career Readiness Certificate Skill Levels

WorkKeys has generated a database with occupational profiles for thousands of jobs across the country. A majority of the jobs require certain skill levels in Reading for Information, Applied Mathematics, and Locating Information. Individuals with higher skill levels qualify for more jobs. The National Career Readiness Certificate uses test results from these assessments to award certificates in three categories:

1. **Bronze Level** signifies an individual has scored at least a Level 3 in each of the three core areas (Reading for Information, Applied Mathematics, and Locating Information) and has the necessary skills for 35% of the jobs in the WorkKeys database.
2. **Silver Level** signifies an individual has scored at least a Level 4 in each of the three core areas and has the necessary skills for 65% of the jobs in the WorkKeys database.
3. **Gold Level** signifies an individual has scored at least a Level 5 in each of the three core areas and has the necessary skills for 90% of the jobs in the WorkKeys database.

Foundations for the Content of the WorkKeys System

ACT established the WorkKeys system in response to a very real need for better information about employability skills and job readiness. To develop the system, ACT consulted with employers, educators, and labor organizations to define essential, foundational workplace skills that are:

- used in a wide range of jobs,
- measurable in large-scale testing settings, and
- teachable in a reasonable period of time.

ACT selected and defined the initial WorkKeys skills based on work with a panel of advisors made up of educators and business persons, reviews of the literature relating to employer-identified skill needs, and a survey of employers and educators. Survey participants, charter members of the WorkKeys development effort, came primarily from seven states and a network of community colleges in California. These charter members and panelists assisted in the design and review of plans and materials, and provided test takers for the prototyping of the system. For more information on the WorkKeys test development process, refer to WorkKeys Technical Manuals for the individual tests or obtain information online at **www.workkeys.com**.

Test users need timely, reliable, and valid information. The WorkKeys tests have been shown to yield reliable measures, and substantial evidence supporting the validity of the tests for their purported uses has been gathered. Some of this evidence of the quality of the tests is summarized in the following sections of this Technical Bulletin. These sections address the reliability, scaling, equating, and validity research for each of the three assessments that make up the National Career Readiness Certificate. In addition, a section on WorkKeys job analysis options is included.

Reliability

For a test to function as intended, the scores need to be reliable and valid. Both of these characteristics have been defined by the *Standards for Educational and Psychological Testing* (1999). Reliability, according to the *Standards,* is "the consistency of . . . measurements when the testing procedures are repeated on a population of individuals or groups." Test publishers are advised to provide reliability indices that reflect random effects on test scores. The indices provided in this chapter fall into three broad categories:

- internal consistency,
- generalizability, and
- classification consistency.

Reliability coefficients are estimates of the consistency of test scores. They range from zero to one, with values near one indicating greater consistency and those near zero indicating little or no consistency.

Internal Consistency for Number-Correct (NC) Scores

Reading for Information and *Applied Mathematics*

Internal consistency reliability measures the consistency within a test by comparing all items with each other. For *Reading for Information* and *Applied Mathematics,* test-data sets were obtained for 121,304 and 122,820 high school students in a Midwestern state in spring 2002 and spring 2003. Internal consistency reliability coefficients for two test forms for each of *Reading for Information* and *Applied Mathematics* were computed. The reliability coefficients (KR-20) for the two forms of the *Reading for Information* test were 0.87 and 0.90, respectively. The reliability coefficients (KR-20) for both forms of the *Applied Mathematics* test were 0.92. These values for reliability coefficients are considered high for a 30-item test.

Locating Information

Internal consistency reliability coefficients for three *Locating Information* forms were computed. Sample sizes for the three forms (Forms 3, 4, and 5) were 3,020, 2,924, and 2,918, respectively. Reliability coefficients (KR-20) were 0.79, 0.83, and 0.79, respectively. These values for reliability coefficients are considered moderately high for a 32-item test.

Generalizability Analyses

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides a broad conceptual and statistical framework for evaluating measurement precision such as reliability. In particular, generalizability theory presents a multidimensional perspective on error variance and enables test users to disentangle multiple sources of error and to estimate the magnitudes of the errors (sampling variabilities). Generalizability analyses produce reliability-like coefficients (*generalizability* and *dependability* coefficients) to indicate reliability of measurement. For example, the univariate generalizability analyses can estimate:

- variability (variance components, σ^2) associated with test takers (p), items (i), and the interaction between test takers and items (pi);
- measurement error variances for norm-referenced (rank-ordering test takers) and domain-referenced (assessing performance level) decisions [$\sigma^2(\delta)$ and $\sigma^2(\Delta)$]; and
- generalizability (reliability-like) coefficients for norm-referenced and domain-referenced decisions ($E\rho^2$ and Φ).

Furthermore, generalizability theory can treat multivariate models in which each test taker has multiple universe scores associated with a specific level of a fixed domain (Brennan, 2001).

A multivariate generalizability theory approach can be used to analyze test data at the level of a table of specifications. For WorkKeys, test items are associated with certain levels of difficulty. In other words, they are nested in levels. Multivariate generalizability analyses can estimate:

- variability associated with items (i) in each fixed level;
- variability associated with item levels (h);
- variability associated with interaction between item levels and test takers (ph);
- generalizability coefficients for the total scores; and
- proportions of the universe score variance at each item level to the variance of the composite (total) universe scores.

Reading for Information

Generalizability analyses (both univariate and multivariate) were conducted using data based on 1,332 test takers. The mean, standard deviation, skewness, and kurtosis of number-correct (NC) scores for these test takers were 20.142, 4.549, –0.628, and 3.269, respectively. Table 1 presents the results of the univariate and multivariate generalizability analyses for *Reading for Information*. The results indicate that items in the middle levels of difficulty contribute most to the universe score variances (the weights). The reliability coefficients, for both rank-ordering test takers and judging test takers' levels of performance, are at or above .80 for the test (see values in bold in Table 1).

Table 1

Estimated Variance Components, Error Variances, and Generalizability Coefficients—*Reading for Information*

Univariate Analysis

Level	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(i)$	$\hat{\sigma}^2(pi)$	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	$\hat{\Phi}$	Weight
3	0.006	0.000	0.026	0.004	0.004	0.582	0.581	0.076
4	0.021	0.000	0.094	0.016	0.016	0.578	0.577	0.193
5	0.048	0.006	0.157	0.026	0.027	0.647	0.639	0.310
6	0.040	0.011	0.200	0.033	0.035	0.548	0.534	0.265
7	0.017	0.002	0.200	0.033	0.034	0.338	0.336	0.156
All Items	0.018	0.061	0.144	0.005	0.007	0.792	0.728	

Multivariate Analysis

	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(h)$	$\hat{\sigma}^2(ph)$	$\hat{\sigma}^2_c(\delta)$	$\hat{\sigma}^2_c(\Delta)$	$E\hat{\rho}^2_c$	$\hat{\Phi}_c$
Total Score	0.019	0.069	0.010	0.005	0.005	**0.804**	**0.800**

Note: Weight indicates the proportional contribution of the universe score variance at each level of items to the composite universe score variance.

Applied Mathematics

Generalizability analyses (both univariate and multivariate) were conducted using data from 1,326 test takers. The mean, standard deviation, skewness, and kurtosis of NC scores were 19.094, 5.765, –0.219, and 2.553, respectively. Table 2 presents the results of the univariate and multivariate generalizability analyses. The results indicate that items in the middle levels of difficulty contribute most to the universe score variances (the weights). The reliability coefficients for both rank-ordering test takers and judging test takers' levels of performance are at or near .88 for the test (see bolded values in Table 2).

Table 2

Estimated Variance Components, Error Variances, and Generalizability Coefficients—*Applied Mathematics*

Univariate Analysis

Level	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(i)$	$\hat{\sigma}^2(pi)$	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	$\hat{\Phi}$	Weight
3	0.018	0.002	0.041	0.007	0.007	0.724	0.714	0.105
4	0.041	0.004	0.103	0.017	0.018	0.707	0.698	0.205
5	0.070	0.003	0.153	0.025	0.026	0.734	0.730	0.283
6	0.057	0.007	0.186	0.031	0.032	0.649	0.641	0.247
7	0.037	0.002	0.170	0.028	0.029	0.564	0.561	0.160
All Items	0.032	0.058	0.143	0.005	0.007	0.871	0.828	

Multivariate Analysis

	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(h)$	$\hat{\sigma}^2(ph)$	$\hat{\sigma}^2_c(\delta)$	$\hat{\sigma}^2_c(\Delta)$	$E\hat{\rho}^2_c$	$\hat{\Phi}_c$
Total Score	0.033	0.065	0.015	0.004	0.004	**0.882**	**0.879**

Note: Weight indicates the proportional contribution of the universe score variance at each level of items to the composite universe score variance.

Locating Information

Generalizability analyses (both univariate and multivariate) were conducted using data based on Form 3 (N = 3,020). The mean, standard deviation, skewness, and kurtosis of NC scores for these test takers were 19.681, 5.401, –0.655, and 0.350, respectively.

Table 3 presents the results of the univariate and multivariate generalizability analyses for Form 3 for *Locating Information*. The results indicate that items in the middle levels of difficulty contribute most to the composite universe score variances (the weight). The reliability coefficients for both rank-ordering test takers and judging test takers' levels of performance are at or near 0.84 for the test (see bolded values in Table 3).

Table 3

Estimated Variance Components, Error Variances, and Generalizability Coefficients—*Locating Information*

Univariate Analysis

Level	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(i)$	$\hat{\sigma}^2(pi)$	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	$\hat{\Phi}$	Weight
3	0.030	0.001	0.055	0.007	0.007	0.812	0.809	0.222
4	0.045	0.009	0.141	0.018	0.019	0.718	0.705	0.326
5	0.038	0.012	0.200	0.025	0.027	0.600	0.586	0.293
6	0.021	0.003	0.176	0.022	0.022	0.485	0.480	0.158
All Items	0.024	0.062	0.153	0.005	0.007	0.833	0.779	

Multivariate Analysis

	$\hat{\sigma}^2(p)$	$\hat{\sigma}^2(h)$	$\hat{\sigma}^2(ph)$	$\hat{\sigma}^2_c(\delta)$	$\hat{\sigma}^2_c(\Delta)$	$E\hat{\rho}^2_c$	$\hat{\Phi}_c$
Total Score	0.024	0.073	0.012	0.004	0.005	**0.843**	**0.837**

Note: Weight indicates the proportional contribution of the universe score variance at each level of items to the composite universe score variance.

Standard Error of Measurement for Number-Correct Scores and Scale Scores

The standard error of measurement (SEM) is closely related to test reliability. The SEM summarizes the amount of error or inconsistency in NC scores on a test. Nonlinear transformations of NC scores to scaled scores alter the relative magnitudes of the conditional SEMs for scaled scores (Kolen, Hanson, & Brennan, 1992). Scale Score average standard errors of measurement were estimated using the 3PL IRT model. The estimated Scale Score reliability (R) was calculated as:

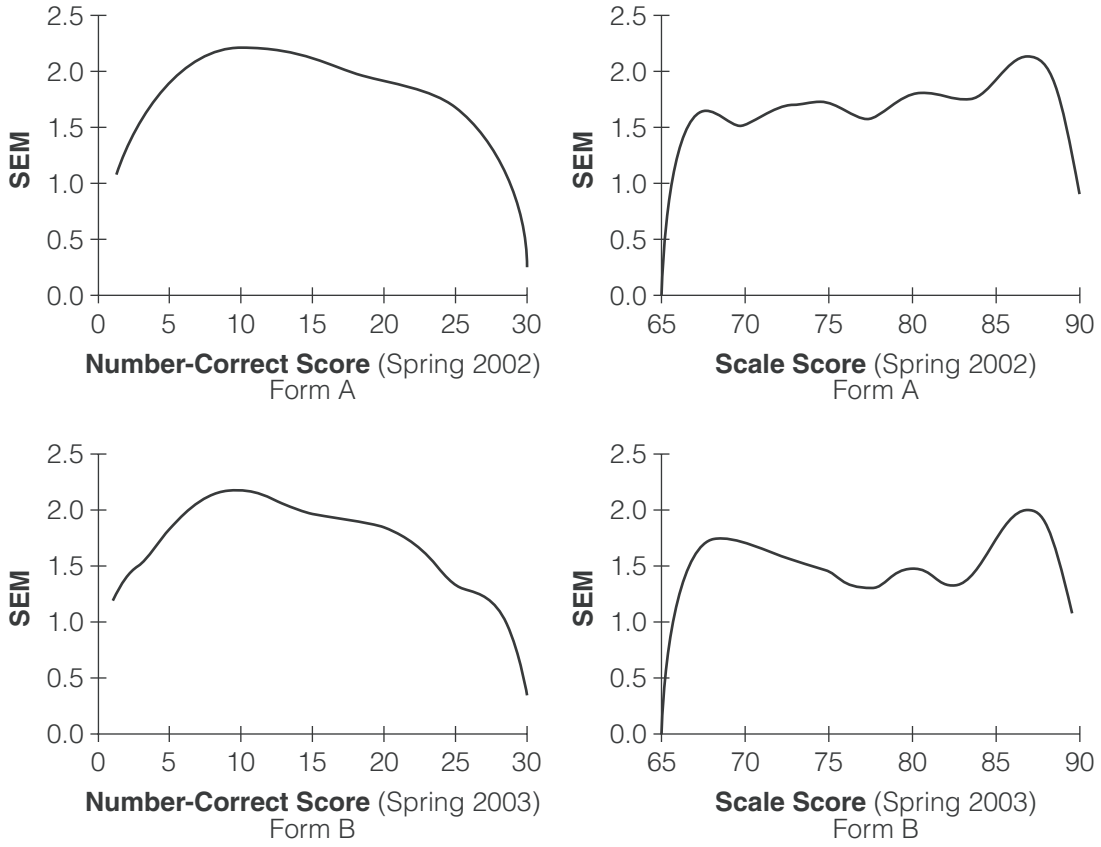
$$R = 1 - \frac{SEM^2}{S^2},$$

where SEM is the estimated Scale Score average standard error of measurement and S is the standard deviation for the observed Scale Scores. This same approach was used for all three tests.

Reading for Information

Figure 2 presents the conditional SEM for two forms of *Reading for Information* as a function of the NC true score (expected NC score), $E(X|\theta)$, and the expected Scale Score, $E(S|\theta)$, based on the 3PL IRT model. The SEMs are generally less than 2 points, showing that the Scale Scores for *Reading for Information* were developed to have approximately constant SEM for all true Scale Scores (i.e., the conditional SEM as a function of true Scale Score is approximately constant).

Figure 2
SEMs for Two Forms of *Reading for Information*

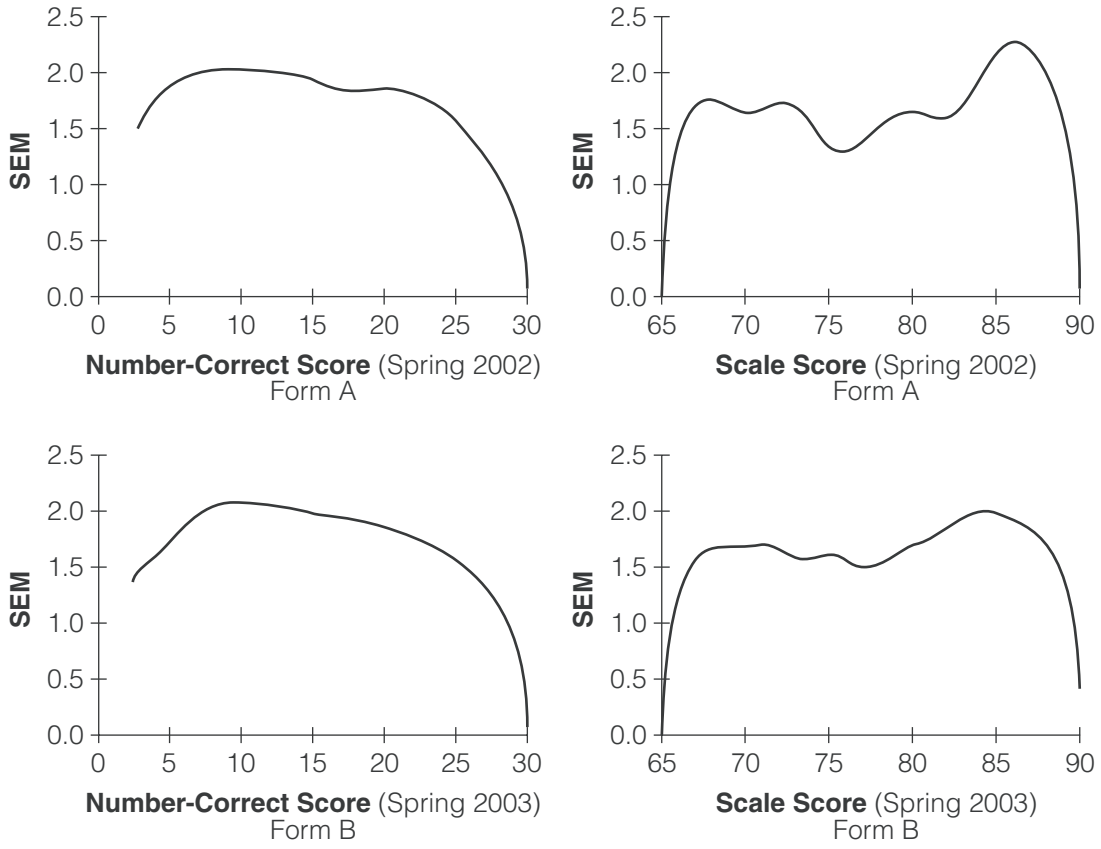


For the Midwestern data sets described above, the Scale Score reliability estimates based on IRT were 0.81 and 0.85. These results are sufficiently high to indicate that test takers' scores should remain fairly constant if test takers repeat the test using alternate forms.

Applied Mathematics

Figure 3 presents the conditional SEM for two forms of *Applied Mathematics* as a function of the NC true score (expected NC score), $E(X|\theta)$, and the expected Scale Score, $E(S|\theta)$, based on the 3PL IRT model. The SEMs are generally less than 2 points, showing that the Scale Scores for this WorkKeys test were developed to have approximately constant SEMs conditional on most Scale Scores.

Figure 3
SEMs for Two Forms of *Applied Mathematics*



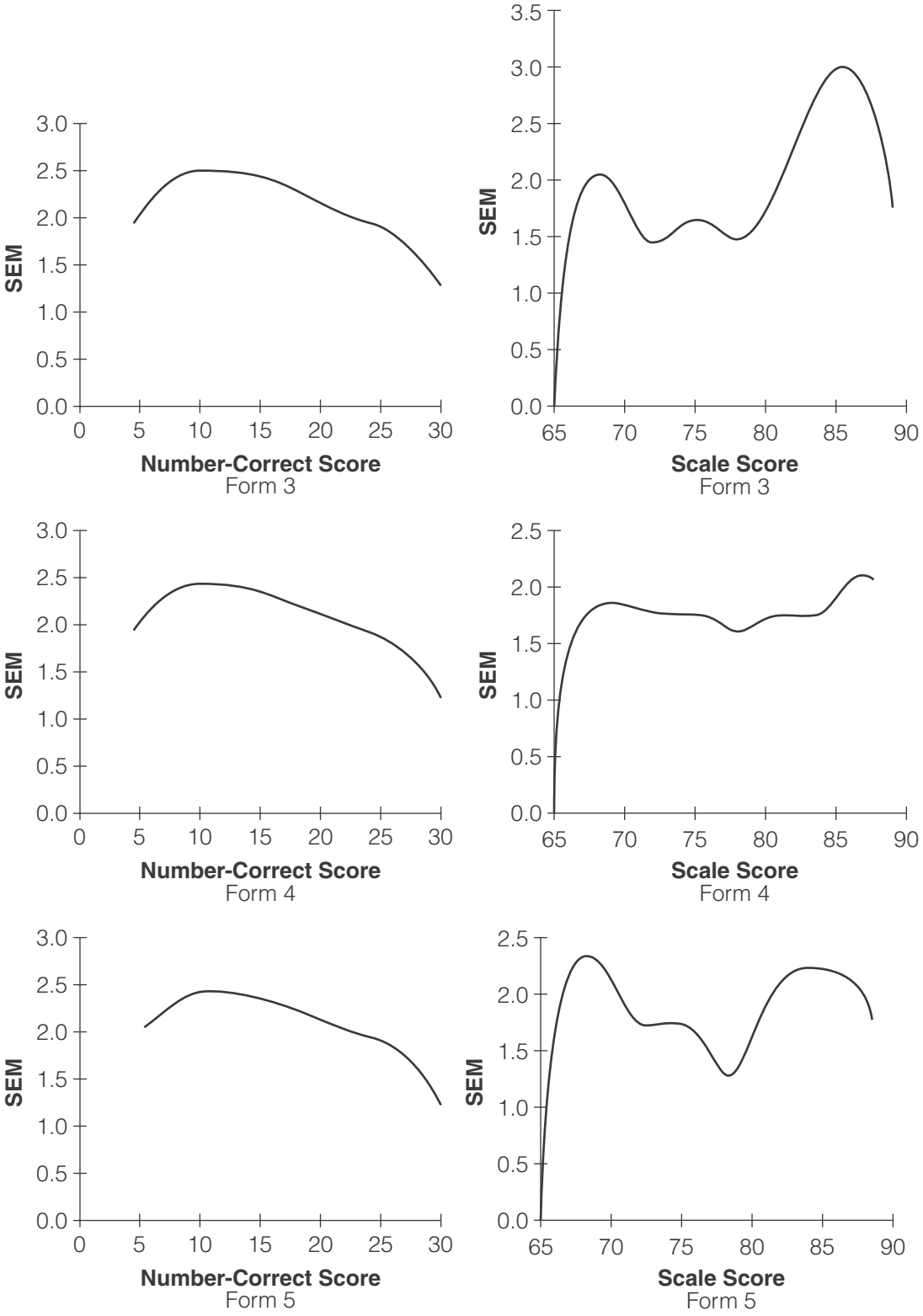
For the Midwestern data sets described above, the Scale Score reliability estimates based on IRT were 0.91 and 0.89. These results are quite consistent across forms, indicating that test takers' scores should remain fairly constant if test takers repeat the test using alternate forms.

Locating Information

Figure 4 presents the conditional SEM for three forms (Forms 3, 4, and 5) of *Locating Information* as a function of the NC true score (expected NC score), $E(X|\theta)$, and the expected Scale Score, $E(S|\theta)$, based on the 3PL IRT model. The SEMs are generally less than 2 points, showing that the Scale Scores for the WorkKeys test were developed to have approximately constant SEM conditional on most Scale Scores.

For the three forms of *Locating Information*, the Scale Score reliability estimates based on IRT were 0.79, 0.79, and 0.82, which indicated that test takers' scores were relatively consistent on the three forms.

Figure 4
SEMs for Three Forms of *Locating Information*



Classification Consistency for Level Scores

WorkKeys tests are often used as *classification* tests. They are designed to permit accurate at-or-above classifications of test takers with regard to the particular level of skill that may be required in a given job setting. Publishers of classification tests are advised to provide information about the percentage of test takers that would be classified in the same way on two applications of the same form or alternative forms (*Standards*, 1999). The *Standards* note that reliability coefficients and standard errors do not directly answer this practical question.

Decision consistency is an important reliability concept for measurements that involve classification decisions and it can help address this practical question. Classification consistency is defined as the extent to which classifications agree when obtained from two independent administrations of a test or two parallel forms of a test. Based on Subkoviak's review (1984), two important classification consistency indices are the agreement index p_o, the proportion of consistent classification based on two parallel forms, and coefficient kappa κ, the proportion of consistent classification adjusted for chance agreement. One principal output from the analysis of classification consistency is a symmetric contingency table. The contingency table can be estimated based on a psychometric model using test scores obtained from a single test administration.

Estimates of classification consistency were derived from a scaling study and pretest data using the IRT methodology described by Schulz, Kolen, and Nicewander (1997, 1999). This methodology performed well when compared with classical methods (Lee, Brennan, & Hanson, 2000). For each test, the 3PL IRT model was fit to the data for the analysis. Indices of classification consistency are more directly informative about the effects of measurement error on a classification test than are SEMs. Classification consistency is defined here as "the proportion or percentage of test takers who would be classified the same way by two parallel tests." As a proportion, classification consistency has the same range as the reliability coefficient: 0 to 1, with 1 being the maximum or best possible. As a percentage, classification consistency ranges from 0 to 100.

Reading for Information

Table 4 shows estimates of classification consistency for *Reading for Information.* The first row, labeled "Exact," shows the percentage of test takers who would receive the same Level Score from two strictly parallel test forms. For example, if a test taker were to take two strictly parallel forms of the test and score at Level 3 on both forms, this would be a case of exact agreement. For *Reading for Information*, it is estimated that such cases would amount to about 55% or 61% of the test takers in this study.

The remaining rows in Table 4 show the consistency of at-or-above classifications separately by level. Entries in the row labeled "≥5," for example, reflect the consistency of classifying test takers with respect to being at or above Level 5. If a test taker were to take two strictly parallel forms of *Reading for Information* and score at Level 4 on the first form and Level 5 on the second, that test taker would not be consistently classified with respect to being at or above Level 5 (≥5), but would be consistently classified with respect to being at or above Level 4 (≥4). Classification consistency is clearly higher for at-or-above classifications than for exact classifications. At-or-above consistency of *Reading for Information* scores is estimated to be not less than 85%, and as high as 98%.

Table 4
Predicted Classification Consistency for Level Scores—*Reading for Information*

	Spring 2002		Spring 2003	
Level	p	kappa	p	kappa
Exact	0.55	0.43	0.61	0.50
≥3	0.97	0.84	0.98	0.89
≥4	0.94	0.76	0.94	0.79
≥5	0.85	0.68	0.87	0.73
≥6	0.85	0.66	0.86	0.63
≥7	0.89	0.55	0.95	0.51

Note: Exact classifications specify a particular skill level for the test taker; "≥" classifications specify that the test taker is at-or-above the indicated level.

Estimates of classification consistency are sensitive to the distribution of skill. For example, the lower boundary on the θ scale for *Reading for Information* Level 5 (0.11) is near zero, which is the mean of the *Reading for Information* θ distribution used to compute classification consistency and classification error. (The θ distribution for each skill is assumed to be standard normal.) This means that the true skill of a relatively large proportion of these test takers was close to the Level 5 boundary. Generally, test takers are more likely to be misclassified because of measurement error when their true skill is closer to the criterion. Given this fact, an 85–87% classification consistency for at-or-above *Reading for Information* Level 5 classification is very good.

By the same reasoning, however, an 89–95% classification consistency for at-or-above *Reading for Information* Level 7 classification is probably overly optimistic. The Level 7 boundary for *Reading for Information*, 2.88, is far above the skill of most test takers in a standard normal θ distribution. Applicants for Level 7 jobs, however, will probably have skills closer to the Level 7 boundary. In that case, the classification consistency for actual job applicants is likely to be lower than is indicated by the values in Table 4 for the *Reading for Information* Level Scores.

Applied Mathematics

Table 5 shows estimates of classification consistency for *Applied Mathematics*. The first row, labeled "Exact," shows the percentage of test takers who would receive the same Level Score from two strictly parallel test forms. For example, if a test taker were to take two strictly parallel forms of the test and scored a Level 3 on both forms, this would be a case of exact agreement. For *Applied Mathematics*, it is estimated that such cases would amount to about 62% of the test takers in this study.

Table 5
Predicted Classification Consistency for Level Scores—*Applied Mathematics*

	Spring 2002		Spring 2003	
Level	p	kappa	p	kappa
Exact	0.63	0.55	0.62	0.54
≥3	0.96	0.79	0.97	0.87
≥4	0.92	0.76	0.93	0.81
≥5	0.92	0.84	0.90	0.80
≥6	0.92	0.83	0.88	0.74
≥7	0.89	0.68	0.93	0.56

Note: Exact classifications specify a particular skill level for the test taker; "≥" classifications specify that the test taker is at-or-above the indicated level.

The remaining rows in Table 5 show the consistency of at-or-above classifications separately by level. Entries in the row labeled "≥5," for example, reflect the consistency of classifying test takers with respect to being at or above Level 5. If a test taker were to take two strictly parallel forms of *Applied Mathematics* and receive a Level Score of 4 on the first form and 5 on the second, he or she would not be consistently classified with respect to being at or above Level 5 (≥5), but would be consistently classified with respect to being at or above Level 4 (≥4). Classification consistency is clearly higher for at-or-above classifications than for exact classifications. At-or-above consistency of *Applied Mathematics* scores are estimated to be not less than 88%, and as high as 97%.

Estimates of classification consistency are sensitive to the skill distribution. For example, the lower boundary on the θ scale for Level 5 of *Applied Mathematics*, 0.36, is near zero, the mean of the *Applied Mathematics* θ distribution used to compute classification consistency and classification error. (The θ distribution for each skill is assumed to be standard normal.) This means that the true skill of a relatively large proportion of these test takers was close to the Level 5 boundary. Generally, the closer a test taker's true level of skill is to a criterion, the more likely he or she is to be misclassified because of measurement error. Given this fact, a 90–92% classification consistency for ≥5 *Applied Mathematics* classification is very good.

By the same reasoning, however, an 89–93% classification consistency for ≥7 classification in *Applied Mathematics* is probably overly optimistic. The Level 7 boundary for *Applied Mathematics*, 2.40, is far above the skill of most test takers in a standard normal θ distribution. Applicants for Level 7 jobs, however, will probably have skills closer to the Level 7 boundary. In that case, the classification consistency for actual job applicants is likely to be lower than the values in Table 5 indicate.

Locating Information

Table 6 shows estimates of classification consistency for three *Locating Information* forms. The first row, labeled "Exact," shows the percentage of test takers who would receive the same Level Score from two strictly parallel test forms. For example, if a test taker were to take two strictly parallel forms of the test and scored a Level 3 on both forms, this would be a case of exact agreement. For *Locating Information*, it is estimated that such cases would be ranged from 60% to 62% of the test takers in this study.

Table 6
Predicted Classification Consistency for Level Scores—*Locating Information*

	Form 3		Form 4		Form 5	
Level	p	kappa	p	kappa	p	kappa
Exact	0.62	0.46	0.60	0.41	0.61	0.44
≥3	0.89	0.64	0.87	0.60	0.85	0.59
≥4	0.80	0.59	0.78	0.56	0.79	0.58
≥5	0.93	0.55	0.92	0.46	0.95	0.53
≥6	1.00	0.18	1.00	0.19	1.00	0.29

Note: Exact classifications specify a particular skill level for the test taker; "≥" classifications specify that the test taker is at-or-above the indicated level.

The remaining rows in Table 6 show the consistency of at-or-above classifications separately by level. Entries in the row labeled "≥5," for example, reflect the consistency of classifying test takers with respect to being at or above Level 5. If a test taker were to take two parallel forms of *Locating Information* and score at Level 4 on the first form and Level 5 on the second, that test taker would not be consistently classified with respect to being at or above Level 5 (≥5), but would be consistently classified with respect to being at or above Level 4 (≥4). Classification consistency is clearly higher for at-or-above classifications than for exact classifications. At-or-above consistency of *Locating Information* scores are estimated to be not less than 78%, and as high as 100%.

By the same reasoning, however, a 100% classification consistency for ≥6 classification in *Locating Information* is probably overly optimistic. The Level 6 boundary for *Locating Information*, 3.48, is far above the skill of most test takers in a standard normal θ distribution. Applicants for Level 6 jobs, however, will probably have skills closer to the Level 6 boundary. In that case, the classification consistency for actual job applicants is likely to be lower than the values in Table 6 indicate.

Scaling and Equating

Level Score Scale

WorkKeys test items are written to assess a certain level of skill applied in a situation involving a certain level of complexity. These levels were initially defined through expert judgment. Pretesting demonstrated that the items met statistical specifications as well. This section describes how the scores based on complete sets of test forms were related back to the same five levels through a process called *scaling*. The equating methods used to establish statistical comparability of the forms are described in the section on *equating*.

The method of assigning Level Scores to test-taker performance was designed to support two basic assumptions.

1. Content experts decided that mastery of a level should mean that a test taker is able to correctly answer 80% of the items representing the level.
2. Test takers have mastery of all levels up to and including the level specified in the score, and do not have mastery of higher levels.

The 80% standard is implemented with respect to a pooled domain of items (not a form-based domain). This pool of items is referred to here as a *level pool* or *level domain*. For *Reading for Information*, each level pool (one each for Level 3 through Level 7) was initially established using eighteen items: six from each of three alternate forms assembled according to the item and test specifications. These three forms had no items in common, but were designed to be comparable in difficulty based on item statistics from pilot studies.

Reading for Information

As there were five level pools and eighteen items in each pool, ninety items were used to define the *Reading for Information* levels. In order to assess mastery using the level pools (18 items), rather than using just the items representing each level on one test form (6 items), an IRT (item response theory) model was used to derive the score scale.

In WorkKeys job analysis, the skill level required for job entry (into the specified job) is established based on the most complex tasks a newly hired employee would be expected to complete using the skill. This remains true even if the job also includes less-complex tasks corresponding to lower levels of the same skill. The WorkKeys scoring system must therefore reflect a reasonable expectation that test takers have mastery of the level specified in the score and mastery of all the preceding levels (Guttman, 1950). For example, a test taker scoring at Level 5 would be expected to have mastered the skills at Levels 5, 4, and 3. However, as multiple-choice test data contain random error, an IRT model was also used to address measurement error.

Scaling Study

The data collection process and the analyses that defined the WorkKeys levels are referred to here as the Level Score *scaling study*. All three test forms were administered to randomly equivalent groups of high school juniors and seniors by spiraling test forms within classrooms. Thus, in each testing room the first person received Form 1, the next person received Form 2, and the next received Form 3; this pattern was repeated so that each form would be given to one-third of the test takers.

Summary statistics for number-correct (NC) scores on the *Reading for Information* forms used in the scaling study are shown in Table 7. Sample sizes for the forms ranged from 2,020 to 2,032.

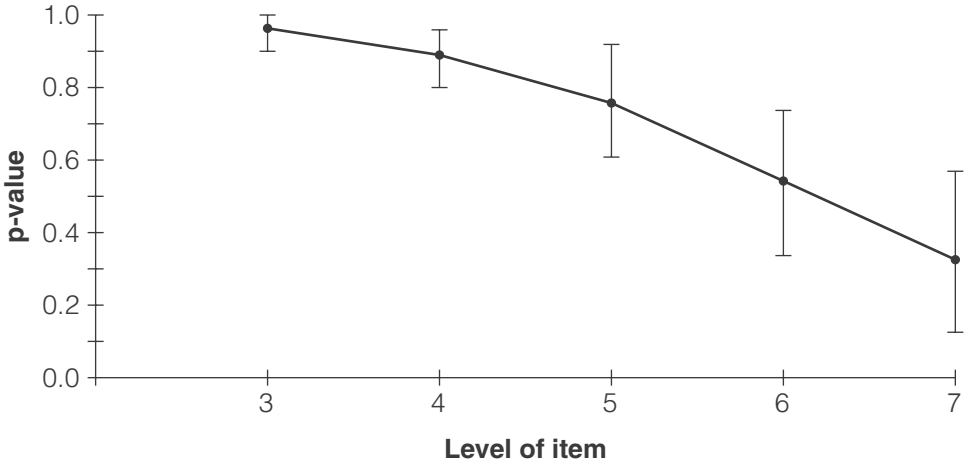
- The mean NC scores ranged from 20.3 to 21.2.
- Skewness (approximately –1) and kurtosis (>1, except for Form 3) were relatively large.
- At .78 to .81, the reliability coefficients based on the three-parameter logistic (3PL) IRT model (Kolen, Zeng, & Hanson, 1996) were similar to the KR-20 reliability coefficients.
- The KR-20 reliability coefficients were .77 to .80 (Schulz, Kolen, & Nicewander, 1999).
- The differences in the coefficients derived by the two methods were not significant.

Table 7
Summary Statistics for *Reading for Information* NC Scores

	Form 1 (N = 2,032)	**Form 2** (N = 2,020)	**Form 3** (N = 2,024)
Mean	20.7	21.2	20.3
Standard Deviation	4.4	4.2	4.5
KR-20	0.79	0.77	0.80
3PL IRT Reliability	0.79	0.78	0.81

Summary statistics for the item p-values from the original *Reading for Information* item-level pools are displayed in Figure 5. This plot shows that while item difficulties overlapped across levels, average item difficulty increased substantially by level (as shown by decreasing mean item p-values).

Figure 5
Item p-Values (p) and Mean Item p-Values (Connected) by Level of Item—*Reading for Information*



The 3PL IRT model was fit to the data separately for each test form using the computer program BILOG (Mislevy & Bock, 1990). Test-taker skill is represented in the 3PL model as a unidimensional, continuous variable, θ (theta). It is assumed that theta is approximately normally distributed in the sample to which the test is administered. Items are represented in the 3PL model by three statistics denoted *a, b,* and *c,* where

a represents the discriminating power of the item,

b represents the difficulty of the item, and

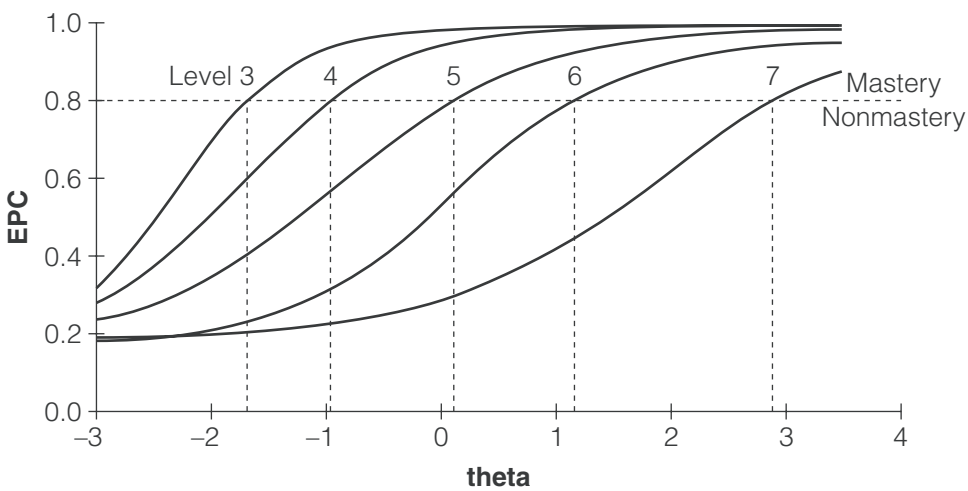
c represents the lower asymptote of the item response function on θ, which is sometimes referred to as the guessing parameter.

The item statistics from the BILOG analyses were used with the IRT model to predict expected proportion correct (EPC) scores on level pools as a function of θ. Figure 6 shows the EPC scores on the *Reading for Information* level pools as a function of the θ for *Reading for Information*:

- The curves in the figure represent level response functions.
- The lower boundary of each level on the θ scale is shown to be the θ coordinate that corresponds to an EPC of 0.8 on the corresponding level pool.

For example, the dotted vertical line on the left in Figure 6 intersects the Level 3 characteristic curve at the coordinates of 0.8 on the EPC axis and at –1.68 on the θ (theta) axis. This means that a test taker with a θ of –1.68 for *Reading for Information* would be expected to select the correct answer for 80% of the items within the Level 3 item pool. Thus, the upper boundary for *Reading for Information* Level 3 is –1.68 on the θ scale.

Figure 6
***Reading for Information* Level Characteristic Curves**



EPC scores represent a test taker's level of skill in two ways that observed scores cannot.

- First, EPC scores represent performance on a larger set of items than those on any given single form. For *Reading for Information*, test takers took only six items per level, but an EPC score represents expected performance on all eighteen items representing the level (six from each of the three forms). EPC scores therefore provide a more consistent basis for assigning Level Scores to test takers who take different forms.

- Second, EPC scores represent levels of performance that do not necessarily correspond to any observed score. In particular, an 80% correct criterion for mastery does not correspond exactly to an NC score for six items (representing a level of *Reading for Information* or *Applied Mathematics* on a single form) or eighteen items (representing the level more generally) and does not correspond exactly to an NC score for eight items (representing a level of *Locating Information* on a form) or sixteen items (representing the level more generally).

The EPC method of defining levels of skill rests on the assumptions that the data fit the IRT model and that the samples of test takers taking alternate forms are randomly equivalent. The fit of the data to the model was evaluated by its ability to predict the observed distributions of Level Scores under three different scoring methods, and to account for observed patterns of mastery over levels (Schulz, Kolen, & Nicewander, 1997 and 1999). The fit of the model for all three tests was judged to be very good in these respects. To estimate the EPC on level pools, item statistics from form-specific BILOG analyses were treated as belonging to a common scale. This treatment rests on the randomly equivalent groups assumption.

Table 8 shows the boundary thetas, form-specific cutoff thetas, and NC score cutoffs that define the levels of *Reading for Information* used in the Level Score scaling study. The lower boundary of Level 3 on the θ scale is shown to be –1.68 (as also illustrated in Figure 6). Similarly, the θ coordinates of the dotted vertical lines representing the lower boundaries of Levels 4, 5, 6, and 7 in Figure 6 are shown in the Lower Boundary column of Table 8 to be –0.95, 0.11, 1.15, and 2.88, respectively.

Table 8
Boundary Thetas, Form Cutoff Thetas, and NC Score Cutoffs—*Reading for Information*

		Form-Specific Cutoff Theta			Number-Correct Score Cutoff		
Level	Lower Boundary	Form 1	Form 2	Form 3	Form 1	Form 2	Form 3
3	–1.68	–1.57	–1.72	–1.66	14	14	13
4	–0.95	–1.04	–1.06	–1.06	17	17	16
5	0.11	0.24	0.13	0.30	22	22	22
6	1.15	1.25	1.02	1.26	25	25	25
7	2.88	2.86	2.73	2.40	28	28	28

Because the theta distribution in a BILOG analysis is assumed to be a standard normal distribution, θ values have approximately the same meaning as Z-scores (standard normal variates) for a distribution of true Level Scores. Such meaning is useful for understanding how difficult it is to achieve a given level of skill. For example, approximately 5% of a standard normal distribution is below a Z-score of –1.68. It is therefore reasonable to suppose that approximately 5% of the test takers who took the *Reading for Information* forms in the scaling study had skills below Level 3.

Table 8 also shows how cutoff scores were selected:

- First, the IRT model was used to find a θ for each NC score on each form.
- The NC score was the true score, rounded to three decimal places (for example, .001), for its corresponding θ (Schulz, Kolen, & Nicewander, 1999).
- The NC score whose θ was the closest to the boundary θ for a level was chosen as the cutoff score for that level.
- The form-specific cutoff θ is the θ corresponding to a cutoff score.

For *Reading for Information* Level 3, the form-specific cutoff thetas were –1.57, –1.72, and –1.66 for Forms 1, 2, and 3, respectively. These thetas were associated with NC scores of 14 for Forms 1 and 2, and 13 for Form 3. On Form 1, the lowest NC score at Level 3 was 14, and the highest NC score at Level 3 was 16. Therefore, for Form 1, NC scores ranging from 14 to 16 were assigned to Level 3.

The fact that the form-specific cutoff thetas do not generally correspond exactly to the boundary thetas reflects the difference between continuous and discrete variables. The EPC and θ scales represent achievement and criterion-referenced standards as continuous variables. These scales can represent a 79% or 81% standard of mastery as precisely as they can an 80% correct standard. NC scores cannot represent all possible standards precisely because they are discrete. For example, a 0.8 EPC has no NC representation in an 18-item level pool.

Across-form variation in the thetas associated with a particular NC score represents a combination of systematic and random effects across forms. Systematic effects include the true psychometric characteristics of the forms. For example, the fact that the θ associated with an NC score of 14 on Form 2 (–1.72) is lower than the θ associated with an NC score of 14 on Form 1 (–1.57) suggests that it may be slightly easier to score at 14 on Form 2 than Form 1. However, random effects (such as the error in estimates of IRT parameters and random differences in ability among test takers in the Form 1 and Form 2 groups) also play a role.

Cutoff scores were often the same across forms. With the exception of the Level 3 and Level 4 cutoff scores for Form 3, the cutoff scores for the *Reading for Information* levels were the same across all forms: 14 for Level 3; 17 for Level 4; 22 for Level 5; 25 for Level 6; and 28 for Level 7. These results attest to the reliability of item statistics from pretest data and to the care taken when these statistics were used to make the alternate forms psychometrically equivalent.

Since the forms were administered to randomly equivalent groups, and cutoff scores were selected to implement standards consistently across forms, the distribution of Level Scores should be similar across forms. Table 9 shows results pertaining to this expectation. The percentage at each level, rounded to the nearest whole number, is shown by form. The percentages at any given level differ by no more than 6 points. These data reflect a fairly even distribution of performance across the sample for all three *Reading for Information* forms.

Table 9
Percentage of Test Takers by Level Scores by Form—*Reading for Information*

Level	Form 1	Form 2	Form 3
Below 3	6	5	6
3	7	7	8
4	38	36	42
5	31	30	27
6	15	19	15
7	2	3	2

The method of selecting cutoff scores is slightly lenient for all three of the tests. The individual-form cutoff is not necessarily higher than the boundary θ. For example, the Level 3 cutoff θ (–1.72) for *Reading for Information* Form 2 is not higher than the Level 3 boundary θ (–1.68). This practice tends to produce a high false-positive-to-false-negative error ratio and a higher overall classification error rate than would occur if the cutoff θ always equaled or exceeded the boundary θ.

A slightly lenient scoring rule was deliberately chosen for two important reasons.

- First, the current scoring procedure replaces one that was also lenient (Schulz, Kolen, & Nicewander, 1997 and 1999). Both the current procedure and the previous one produce similar frequency distributions of observed Level Scores. This is important for connecting current results with past results for WorkKeys users.
- Second, a lenient implementation of the 0.8 EPC standard in WorkKeys is justified by the error inherent in measuring with reference to a standard.

In addition to the measurement error associated with a test taker's score, there is also error in setting a criterion-referenced standard. One or both of these types of errors are typically cited in choosing a cutoff score that is more lenient, and gives the benefit of the doubt to the test taker.

Leniency typically takes the form of a cutoff score that is one or more standard errors of measurement below the score that strictly represents the standard. ACT's particular method of scoring WorkKeys tests is less lenient than this. Strict implementation of the 0.8 EPC standard would require the cutoff θ to exceed the boundary θ. In about half the cases, it already does. In the other half, the cutoff score would be a lower value than would be required by a strict implementation of the standard. One NC point of difference is less than one standard error of measurement on the NC scale for the WorkKeys tests.

Applied Mathematics

The Level Scoring method for the WorkKeys *Applied Mathematics* assessment was developed using the data from three alternate forms. These three forms had no items in common, but were designed to be comparable in difficulty based on item statistics from pretest studies. There are five skill levels for *Applied Mathematics* (Level 3 to Level 7), and each is represented by six items on each form. Thus there are eighteen items per level, thirty items per form, and a total of ninety items that were used to define the *Applied Mathematics* levels.

The data collection process and the analyses that defined the WorkKeys levels are referred to here as the Level Score *scaling study*. All three test forms were administered to randomly equivalent groups of high school juniors and seniors by spiraling test forms within classrooms. Thus, in each testing room the first person received Form 1, the next person received Form 2, and the next received Form 3; this pattern was repeated so that each form would be given to one-third of the test takers.

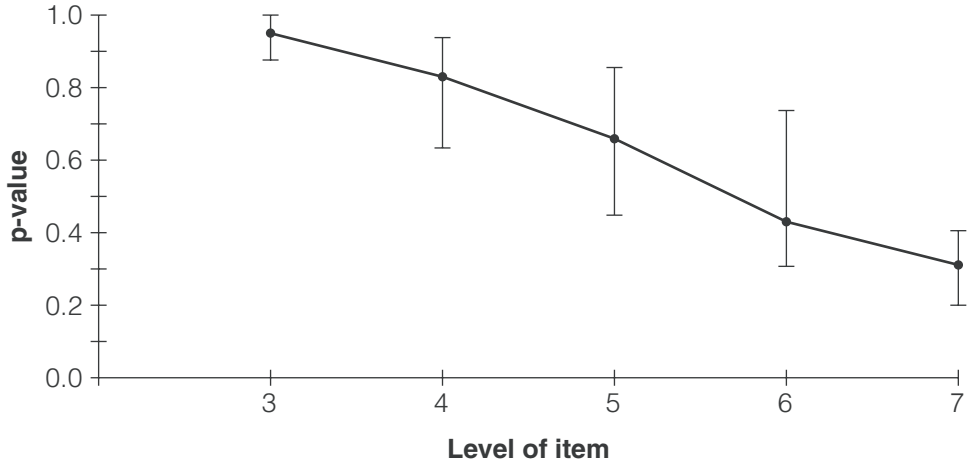
Summary statistics for number-correct (NC) scores on the *Applied Mathematics* forms used in the scaling study are shown in Table 10. Sample sizes for the forms ranged from 1,996 to 2,046. The mean NC scores ranged from 18.8 to 19.1. Reliability coefficients based on the three-parameter logistic (3PL) IRT model (Kolen, Zeng, & Hanson, 1996) were slightly higher (.82 to .85) than the KR-20 reliability coefficients (.80 to .83) (Schulz, Kolen, & Nicewander, 1999).

Table 10
Summary Statistics for *Applied Mathematics* NC Scores

	Form 1 (N = 2,022)	**Form 2** (N = 2,046)	**Form 3** (N = 1,996)
Mean	18.8	19.0	19.1
Standard Deviation	5.1	4.9	4.8
KR-20	0.83	0.81	0.80
3PL IRT Reliability	0.85	0.83	0.82

The p-values of the items comprising the original *Applied Mathematics* level pools are displayed in Figure 7. This plot shows that while difficulties overlapped across levels, average item difficulty increased substantially by level (as shown by decreasing mean item p-values).

Figure 7
Item p-Values (p) and Mean Item p-Values (Connected) by Level of Item—*Applied Mathematics*



The 3PL IRT model was fit to the data separately for each test form using the computer program BILOG (Mislevy & Bock, 1990).

The item statistics from the BILOG analyses were used with the IRT model to predict expected proportion correct (EPC) scores on level pools as a function of θ. Figure 8 shows the EPC scores on *Applied Mathematics* level pools as a function of *Applied Mathematics* θ (theta). The curves in this figure represent level response functions. The lower boundary of each *Applied Mathematics* level on the θ scale is shown to be the θ coordinate corresponding to an EPC of 0.8 on the corresponding level pool. For example, the vertical dotted line on the left in Figure 8 intersects the Level 3 characteristic curve at the coordinates of 0.8 on the EPC axis and at –1.43 on the θ (theta) axis. This means that a test taker with an *Applied Mathematics* θ of –1.43 would be expected to get 80% of the items correct within the Level 3 item pool. The boundary for *Applied Mathematics* Level 3 is thus –1.43 on the θ scale.

Figure 8
***Applied Mathematics* Level Characteristic Curves**

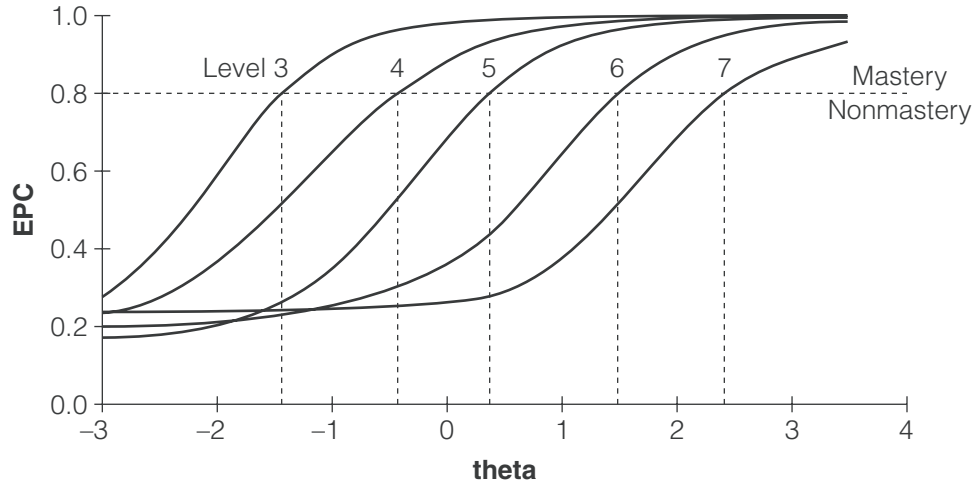


Table 11 shows the boundary thetas, form-specific cutoff thetas, and NC score cutoffs that define the levels of *Applied Mathematics* used in the Level Score scaling study. The lower boundary of Level 3 on the θ scale for *Applied Mathematics* is shown to be –1.43, as illustrated in Figure 8. Similarly, the θ coordinates of the dotted vertical lines representing the lower boundaries of Levels 4, 5, 6, and 7 in Figure 8 are shown in the Lower Boundary column of Table 11 to be, respectively, –0.43, 0.36, 1.48, and 2.40.

Table 11
Boundary Thetas, Form Cutoff Thetas, and NC Score Cutoffs—*Applied Mathematics*

		Form-Specific Cutoff Theta			Number-Correct Score Cutoff		
Level	Lower Boundary	Form 1	Form 2	Form 3	Form 1	Form 2	Form 3
3	–1.43	–1.43	–1.51	–1.54	12	12	12
4	–0.43	–0.37	–0.47	–0.49	17	17	17
5	0.36	0.48	0.42	0.40	21	21	21
6	1.48	1.28	1.36	1.36	25	25	25
7	2.40	2.34	2.19	2.56	29	28	28

Because the θ distribution in a BILOG analysis is assumed to be a standard normal distribution, θ values have approximately the same meaning as Z-scores (standard normal variates) for distribution of true Level Scores. Such a meaning is useful for understanding how difficult it is to achieve a given level of skill. For example, approximately 8% of a standard normal distribution is below a Z-score of –1.43. It is therefore reasonable to suppose that approximately 8% of the test takers who took the *Applied Mathematics* forms in the scaling study had skills below Level 3.

Table 11 also shows how cutoff scores were selected. First, the IRT model was used to find a θ for each NC score on each form. The NC score was the true score, rounded to three decimal places (for example, .001), for its corresponding θ (Schulz, Kolen, & Nicewander, 1999). The NC score whose θ was the closest to the boundary θ for a level was chosen as the cutoff score for that level. The form-specific cutoff θ is the θ corresponding to a cutoff score. For *Applied Mathematics* Level 3 (as shown in Table 11), the form-specific cutoff θs were –1.43, –1.51, and –1.54 for Forms 1, 2, and 3, respectively. These θs were associated with an NC score of 12 across all three forms. On Form 1, the lowest NC score at Level 3 was 12 and the highest NC score at Level 3 was 16. Therefore, the NC scores ranging from 12 to 16 were assigned to Level 3. The fact that the form-specific cutoff thetas do not generally correspond exactly to the boundary thetas reflects the difference between continuous and discrete variables. The EPC and θ scales represent achievement and criterion-referenced standards as continuous variables. These scales can represent a 79% or 81% standard of mastery as precisely as an 80% correct standard. NC scores cannot represent all possible standards precisely because they are discrete. For example, a 0.8 EPC has no NC representation in an 18-item level pool.

Across-form variation in the θs associated with a particular NC score represents a combination of systematic and random effects across forms. Systematic effects include the true psychometric characteristics of the forms. For example, the fact that the θ associated with a 12 on Form 2 (–1.51) is lower than the θ associated with a 12 on Form 1 (–1.43) suggests that it may be slightly easier to get a 12 on Form 2 than on Form 1. However, random effects (such as the error in estimates of IRT parameters and random differences in ability among test takers in the Form 1 and Form 2 groups) also play a role. Remarkably, cutoff scores were often the same across forms. These results attest to the reliability of item statistics from pilot data and to the care with which these statistics were used to make the alternate forms as equivalent as possible.

Since the forms were administered to randomly equivalent groups, and cutoff scores were selected to implement standards consistently across forms, the distributions of Level Scores should be similar across forms. Table 12 shows results pertaining to this expectation. The percentage at each level of *Applied Mathematics*, rounded to the nearest whole number, is shown by form. The percentages at a given level differ by no more than 4 points.

Table 12
Percentage of Test Takers by Level Scores by Form—*Applied Mathematics*

Level	Form 1	Form 2	Form 3
Below 3	8	7	7
3	22	20	20
4	31	32	32
5	25	28	29
6	10	9	9
7	2	3	2

As previously stated, the method of selecting cutoff scores is slightly lenient. The individual-form cutoff is not necessarily higher than the boundary θ. For example, the Level 3 cutoff θ for Form 2 (–1.51) is not higher than the Level 3 boundary θ of –1.43. This practice tends to produce a high false-positive-to-false-negative error ratio and a higher overall classification error rate than would occur if the cutoff θ always equaled or exceeded the boundary θ.

Locating Information

The Level Scoring method for the WorkKeys *Locating Information* assessment was developed using the data from two alternate forms. These two forms had no items in common, but were designed to be comparable in difficulty based on item statistics from pilot studies. There are four skill levels for *Locating Information* (Level 3 to Level 6), and each level was represented by eight items on each form (16 items per level pool). Thus there were sixty-four items ($16 \times 4 = 64$) in total for defining the *Locating Information* levels.

The data collection process and the analyses that defined the WorkKeys levels are referred to here as the Level Score *scaling study*. Two test forms were administered to randomly equivalent groups of high school juniors and seniors by spiraling test forms within classrooms. Thus, in each testing room the first person received Form 1, and the next received Form 2; this pattern was repeated so that each form would be given to half of the test takers.

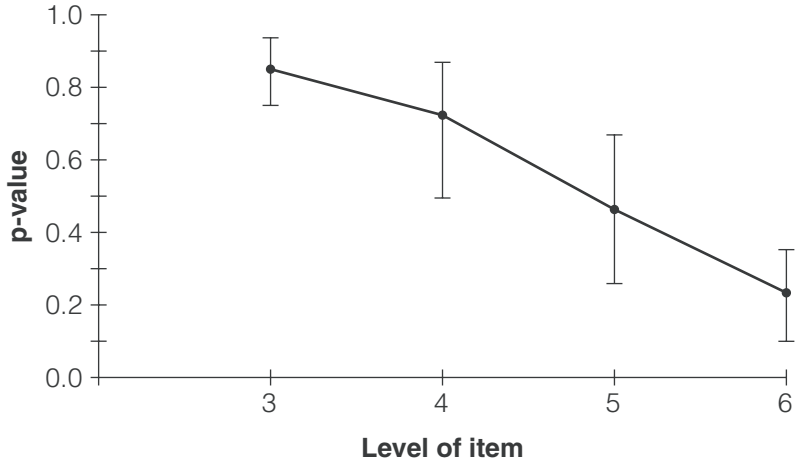
Summary statistics for number-correct (NC) scores on the *Locating Information* forms used in the scaling study are shown in Table 13. Sample sizes for the forms are 1,321 and 1,300. The mean NC scores were 18.46 and 18.05, respectively. Reliability coefficients based on the three-parameter logistic (3PL) IRT model (Kolen, Zeng, & Hanson, 1996) were slightly higher (.82 and .81) than were the KR-20 reliability coefficients (.81 and .79), respectively. Note that the reliability coefficient for 3PL is based on item response theory, and the KR-20 reliability coefficient is based on classical test theory.

Table 13
Summary Statistics for *Locating Information* NC Scores

	Form 1 (N = 1,321)	**Form 2** (N = 1,300)
Mean	18.46	18.05
Standard Deviation	5.21	4.99
KR-20	0.81	0.79
3PL IRT Reliability	0.82	0.81

The p-values of the items comprising the original *Locating Information* level pools are displayed in Figure 9. This plot shows that while item difficulties overlapped across levels, average item difficulty increased substantially by level (as shown by decreasing mean item p-values).

Figure 9
Item p-Values (p) and Mean Item p-Values (Connected) by Level of Item—*Locating Information*



The 3PL IRT model was fit to the data separately for each test form using the computer program BILOG (Mislevy & Bock, 1990). The item statistics from the BILOG analyses were used with the IRT model to predict expected proportion

correct (EPC) scores on level pools as a function of θ. Figure 10 shows the EPC scores on *Locating Information* level pools as a function of *Locating Information* θ (theta). The curves in this figure represent level response functions. The lower boundary of each *Locating Information* level on the θ scale is shown to be the θ coordinate corresponding to an EPC of 0.8 on the corresponding level pool. For example, the vertical dotted line on the left in Figure 10 intersects the Level 3 characteristic curve at the coordinates of 0.8 on the EPC axis and at –0.75 on the θ (theta) axis. This means that a test taker with a *Locating Information* θ of –0.75 would be expected to get 80% of the items correct within the Level 3 item pool. The boundary for *Locating Information* Level 3 is thus –0.75 on the θ scale.

Figure 10
***Locating Information* Level Characteristic Curves**

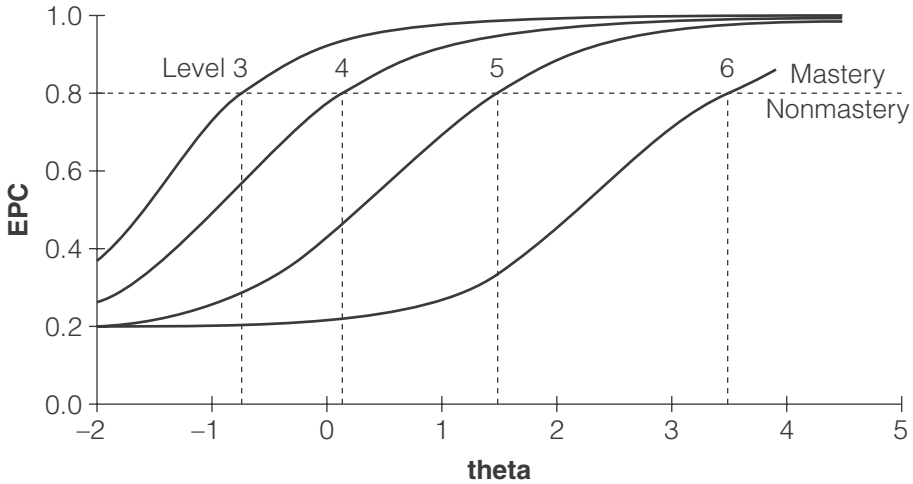


Table 14 shows the boundary thetas, form-specific cutoff thetas, and NC score cutoffs that define the levels of *Locating Information* used in the Level Score scaling study. The lower boundary of Level 3 on the θ scale for *Locating Information* is shown to be –0.75, as illustrated in Figure 10. Similarly, the θ coordinates of the dotted vertical lines representing the lower boundaries of Levels 4, 5, and 6 in Figure 10 are shown in the Lower Boundary column of Table 14 to be, respectively, 0.13, 1.48, and 3.48.

Table 14
Boundary Thetas, Form Cutoff Thetas, and NC Score Cutoffs—*Locating Information*

		Form-Specific Cutoff Theta		Number-Correct Score Cutoff	
Level	Lower Boundary	Form 1	Form 2	Form 1	Form 2
3	–0.75	–0.74	–0.71	15	15
4	0.13	0.06	0.08	19	19
5	1.48	1.47	1.51	25	24
6	3.48	3.57	3.33	31	29

Because the θ distribution in a BILOG analysis is assumed to be a standard normal distribution, θ values have approximately the same meaning as Z-scores (standard normal variates) for distribution of true Level Scores. Such meaning is useful for understanding how difficult it is to achieve a given level of skill. For example, approximately 23% of a standard normal distribution is below a Z-score of –0.75. As mentioned earlier, the boundary for *Locating Information* Level 3 is –0.75, which means approximately 23% of the test takers who took the *Locating Information* forms in the scaling study had skills below Level 3.

Table 14 also shows how cutoff scores were selected. First, the IRT model was used to find a θ for each NC score on each form. The NC score was the true score, rounded to three decimal places (for example, 0.001), for its corresponding θ (Schulz, Kolen, & Nicewander, 1999). The NC score whose θ was the closest to the boundary θ for a level was chosen as the cutoff score for that level. The form-specific cutoff θ is the θ corresponding to a cutoff score. For *Locating Information* Level 3 (as shown in Table 14), the form-specific cutoff thetas were –0.74 and –0.71 for Forms 1 and 2, respectively. These thetas were associated with NC scores of 15 for Forms 1 and 2. On Form 1, the lowest NC score at Level 3 was 15, and the highest NC score at Level 3 was 18. Therefore, for Forms 1 and 2, NC scores ranging from 15 to 18 were assigned to Level 3.

The fact that the form-specific cutoff thetas do not generally correspond exactly to the boundary thetas reflects the difference between continuous and discrete variables. The EPC and θ scales represent achievement and criterion-referenced standards as continuous variables. These scales can represent a 79% or 81% standard of mastery as precisely as an 80% correct standard. NC scores cannot represent all possible standards precisely because they are discrete. For example, a 0.8 EPC has no NC representation in a 16-item level pool. Across-form variation in the thetas associated with a particular NC score represents a combination of systematic and random effects across forms. Systematic effects include the true psychometric characteristics of the forms. For example, the fact that the θ associated with a 15 on Form 1 (–0.74) is lower than the θ associated with a 15 on Form 2 (–0.71) suggests that it may be slightly easier to get a 15 on Form 1 than on Form 2. However, random effects (such as the error in estimates of IRT parameters and random differences in ability among test takers in the Form 1 and Form 2 groups) also play a role. From Table 14, cutoff scores were often the same across forms. These results attest to the reliability of item statistics from pretest data and to the care with which these statistics were used to make the alternate forms as equivalent as possible.

Since the forms were administered to randomly equivalent groups, and cutoff scores were selected to implement standards consistently across forms, the distributions of Level Scores should be similar across forms. Table 15 shows results pertaining to this expectation. The percentage at each level of *Locating Information*, rounded to the nearest integer, is shown by form. The percentages at any given level differ by no more than 4 percentage points.

Table 15
Percentage of Test Takers by Level Scores by Form—*Locating Information*

Level	Form 1	Form 2
Below 3	20	20
3	24	26
4	48	44
5	9	10
6	0	0

Again, the method of selecting cutoff scores is slightly lenient. The individual-form cutoff is not necessarily higher than the boundary θ. For example, the Level 4 cutoff θ (0.06) for Form 1 is lower than the Level 4 boundary θ (0.13).

Scale Scores

Scaling is a process of setting up a rule of correspondence between the observed scores and the numbers assigned to them. The usefulness of a score scale depends on whether or not it can facilitate meaningful interpretation and can minimize misinterpretation and unwarranted inferences (Petersen, Kolen, & Hoover, 1989). The purpose of developing an additional score scale for each WorkKeys test was to provide users with more detailed information for use in program evaluation and outcome measurement. Therefore, the new score scale makes finer distinctions than can be made with the Level Score scale.

The Scale Scores for the WorkKeys tests were developed using the equal standard error of measurement methodology developed by Kolen (1988). First, the Number Correct (NC) scores were transformed using the arcsine transformation described by Freeman and Tukey (1950) to stabilize error variance. The form of this transformation is

$$c(i) = \frac{1}{2}\left(\sin^{-1}\sqrt{\frac{i}{K+1}} + \sin^{-1}\sqrt{\frac{i+1}{K+1}}\right)$$

where sin^{-1} is the arcsine function and K is the number of items. This nonlinear transformation is designed to equalize error variance across the score points. The transformed arcsine values were then linearly transformed to the new score scale using

$$s = A^{*}c(i) + B$$

where s is the Scale Score, A is the slope and B is the intercept. More specifically,

$$A = (s_1 - s_2)/[c(i_1) - c(i_2)] \text{ and } B = s_2 - A^{*}c(i_2) \text{ or } B = s_1 - A^{*}c(i_1),$$

where s_1 and s_2 correspond to the lowest and highest Scale Score points, respectively. The non-integer Scale Scores were rounded to integers to obtain reported scores.

A 25-point scale (65–90) was chosen for the linear transformation after various other lengths of scale had been considered. In consideration of the "guessing effect," scores at the lower end were truncated. A combination of classical test theory and IRT was used to determine at what score truncation should occur. The goals were to:

- provide an adequate number of score points for the anticipated uses of the scores and to
- avoid using more score points than the number of items could support.

For practical reasons, the score scale for a particular test form may also need to be adjusted to preserve specified score scale ranges, to prevent large gaps in the score scale, to avoid having too many NC scores converting to a single Scale Score, and especially to preserve a one-to-one mapping between a cutoff score for a level and a Scale Score. This procedure resulted in the final conversions from NC scores to Scale Scores.

The Scale Score is a function of the NC score. Scale Scores also incorporate equal conditional standard error of measurement (SEM) along most of the score scale. The standard error of measurement was about 1.5 to 2 points, so an approximate 68% confidence interval could be formed by adding ±2 points to test takers' Scale Scores. After the score scale was developed for a base form, equating was performed to obtain score scale conversions for all other forms. Two designs were used for equating:

- randomly equivalent groups and
- common-item nonequivalent groups.

Pre-equating methods based on IRT can also be used for future forms. Summary statistics for the distributions of NC scores and Scale Scores for the test takers in a statewide testing program for *Reading for Information* and *Applied Mathematics* grade 11 test takers are shown in Tables 16 and 17 and summary statistics for the distributions of NC scores and Scale Scores for three *Locating Information* forms are shown in Table 18.

Table 16
Summary Statistics for *Reading for Information* NC Scores and Scale Scores

	N	Score	Mean	Standard Deviation
Spring 2002	121,304	NC Score	21.75	4.42
		Scale Score	79.15	3.87
Spring 2003	122,820	NC Score	21.03	4.49
		Scale Score	78.73	3.68

Table 17
Summary Statistics for *Applied Mathematics* NC Scores and Scale Scores

	N	Score	Mean	Standard Deviation
Spring 2002	121,304	NC Score	22.12	5.83
		Scale Score	79.18	5.58
Spring 2003	122,820	NC Score	21.29	5.48
		Scale Score	79.05	5.15

Table 18
Summary Statistics for *Locating Information* NC Scores and Scale Scores

	N	Score	Mean	Standard Deviation
Form 3	3,020	NC Score	19.68	5.40
		Scale Score	74.67	3.84
Form 4	2,924	NC Score	18.27	4.86
		Scale Score	74.72	3.76
Form 5	2,918	NC Score	17.95	4.95
		Scale Score	74.58	3.76

Equating

New forms of the WorkKeys tests are developed as needed. Though each form is constructed to adhere to the same content and statistical specifications, the forms may be slightly different in difficulty. To control for these differences, scores on all forms are equated so that when they are reported to test takers (as either Level Scores or Scale Scores), equated Scale Scores have the same meaning regardless of the particular form administered. Thus, Level Scores and Scale Scores are comparable across test forms and test dates. However, they do *not* compare across tests. A Level Score of 3 or a Scale Score of 65 in *Reading for Information* does not compare in any way to a Level Score of 3 or a Scale Score of 65 on any other WorkKeys test. Two common equating designs are used with the WorkKeys tests (Kolen & Brennan, 1995).

In a *randomly equivalent groups design*, new test forms are administered along with an anchor form that has already been equated to previous forms. A spiraling process is used to distribute test forms to test takers. Thus, in each testing room the first person receives Form 1, the next Form 2, and the next Form 3. This pattern is repeated so that each form is given to one-third of the test takers and the forms are given to randomly equivalent groups. When this design is used, the difference in total-group performance on the new and anchor forms is considered a direct indication of the difference in difficulty between the forms. Scores on the new forms are equated to the score scale using various equating methodologies including linear and equipercentile procedures (e.g., see Kolen & Brennan, 1995). When the Level Score and Scale Score conversions are chosen for each form, the equating functions are examined, as are the resulting distributions of the scores and their means, standard deviations, skewness, and kurtosis.

A *common-item nonequivalent groups design* has been used when a spiraling technique cannot be implemented in a test administration, when only a single form can be administered per test date, or when some items are changed in a revised form. In a common-item nonequivalent groups design, the new form(s) and base form have a set of items in common. These common items (anchors) are chosen to represent the content and statistical characteristics of the test and are usually interspersed among the other items in the new test form(s). The different forms are then administered to different groups of test takers. In this design, the groups are not assumed to be equivalent. The common items are used to adjust for group differences. Observed differences between group performances can result from a combination of (a) test taker group differences and (b) test form differences. Strong statistical assumptions are usually required to separate these differences.

The various equating methods under the common-item nonequivalent groups design are distinguished in terms of their statistical assumptions (e.g., see Kolen & Brennan, 1995). *Observed-score equating* methods are usually used in equating WorkKeys test forms. For each form, the equating functions are examined, as are the resulting distributions of scaled scores and the mean, standard deviation, skewness, and kurtosis of the scaled scores. The set of equating conversions chosen for each form is the one that results in scaled score distributions and scaled score moments that are judged to be reasonable based on the sample sizes, the magnitudes of the form differences and group differences, and the historical statistics for the test.

Pool Calibration and Pre-equating

After being field tested and reviewed, the operational items and pretest items are calibrated and placed in the item pool. The initial item pool was developed using the items available at that time, and new items are added to it as they are developed. To calibrate the item pool, a linking plan was developed. The plan listed all the forms and their links to a base form. A 3PL IRT model was used in the calibration. Items were calibrated either concurrently or separately using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), depending on the data-collection designs. Items in the different forms administered under the randomly equivalent groups design were calibrated concurrently. The item parameter estimates for all the forms administered were placed on the scale for the base form using the Stocking-Lord characteristic curve method (Stocking & Lord, 1983).

During the calibration process, item statistics from both classical test theory and IRT analyses are reviewed. Pretest items with very low discrimination indices are excluded from the pool. All of the estimated item parameters from multiple calibrations of any set of items are plotted and compared to each other. For most items, the estimates are similar. If they are not, the item is not used as an anchor item in the scaling process. The estimates for pretest items are replaced by the estimates from operational test administrations when they become available.

A calibrated item pool is a group of items that have their item parameter estimates placed on a common scale. Creating an IRT-calibrated item pool makes it possible to pre-equate new forms prior to actual test administration. The item parameter estimates can be used in assembling new forms and conducting pre-equating. As described above, most WorkKeys forms are currently equated using either randomly equivalent groups or common-item equating methods. When these two conventional equating designs cannot be implemented, pre-equating is performed if the items have been calibrated previously. When possible, research studies are carried out to evaluate the comparability of (a) the pre-equating results and (b) the equating results derived from other methods (Gao, Harris, Yi, & Lei, 2003). In addition, the stability of parameter estimates of pretest items and their impact on pre-equating are evaluated (Gao, Chen, & Harris, 2005).

WorkKeys Validity Evidence

The *Uniform Guidelines on Employee Selection Procedures* (1978) describes three kinds of validity: content validity, construct validity, and criterion-related validity. At the same time, the *Standards for Educational and Psychological Testing* (1999) describes validity as a unitary concept supported by three kinds of evidence. That evidence can be "based on relations to other variables" (criterion-related), "based on test content" (content validity), or established by "the validity argument" (construct-related), which is "an explicit scientific justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores" (pp. 13, 11, 184, 174). Thus, evidence may be accumulated in a number of ways. What is relevant is that validity is established as a whole. The value of each way of collecting evidence is determined by its appropriateness to the situation, not by any inherent value of its own.

The *Standards* (pp. 9–11) also explains that the need for validity evidence is based on the assumption that a test is going to be used for a purpose, and it is necessary to provide evidence showing that using the test for that purpose is appropriate. When a test is administered, and a score is reported, it is necessary to determine what the score means within the context in which the test is being used. This is done by accumulating evidence to show what the test score means. Thus, validity refers to the degree to which the evidence supports the interpretation of the scores. This is what is validated, not the test itself.

The WorkKeys assessments are designed for use in both business and educational settings. To support these uses, validity evidence must be obtained in several different contexts. The validation process begins with the statement of what the score is expected to indicate and an explanation of how it will make that indication. Validation is achieved when a scientifically sound validity argument has been presented. Such an argument supports the intended interpretation of the test scores by showing what they mean within a specific context. The validation process is unitary in that when the evidence is collected and analyzed, the results will be described in terms of one concept, "validation." The WorkKeys system relies primarily on content validation; criterion-related and construct-related data are collected when appropriate and necessary.

Construct-Related Evidence

Construct-related evidence for test validity focuses primarily on the test score as a measure of the psychological characteristic of interest. The process of compiling construct-related evidence starts with test development and continues until the pattern of empirical relationships between test scores and other variables clearly indicates the meaning of the test score.

WorkKeys *Reading for Information* and the ACT Assessment®

One would expect scores on WorkKeys *Reading for Information,* which focuses on the reading and understanding of work-related instructions and policies, to be related to other reading test scores. The relationship between WorkKeys *Reading for Information* and the ACT Reading and English Tests provides construct evidence for test validity. The ACT Reading and English Tests are part of the ACT test, which is designed to measure the skills acquired during secondary education that are most important to success in postsecondary education. The material the tests cover emphasizes the major content areas that are prerequisite to successful performance in entry-level courses in college reading and English.

The results listed in Tables 19 and 20 and in Figure 11 (based on testing in a Midwestern state in 2002, n = 121,304 and 2003, n = 122,820) show that there is a moderate relationship between WorkKeys *Reading for Information* scores and scores on the ACT Reading test. In general, test takers who received higher Level Scores on *Reading for Information* also received higher Scale Scores on ACT Reading.

Table 19

Correlations between WorkKeys *Reading for Information,* ACT Reading, and ACT English

		NC Score		Scale Score	
		ACT Reading	**ACT English**	**ACT Reading**	**ACT English**
	Range	1–40	1–75	1–36	1–36
2002	WorkKeys *Reading for Information*	0.650	0.692	0.608	0.639
	ACT Reading		0.807		0.812
2003	WorkKeys *Reading for Information*	0.657	0.711	0.620	0.660
	ACT Reading		0.791		0.799

Table 20

Percents of Test Takers by WorkKeys *Reading for Information* Level Scores and Ranges of ACT Reading Scale Scores

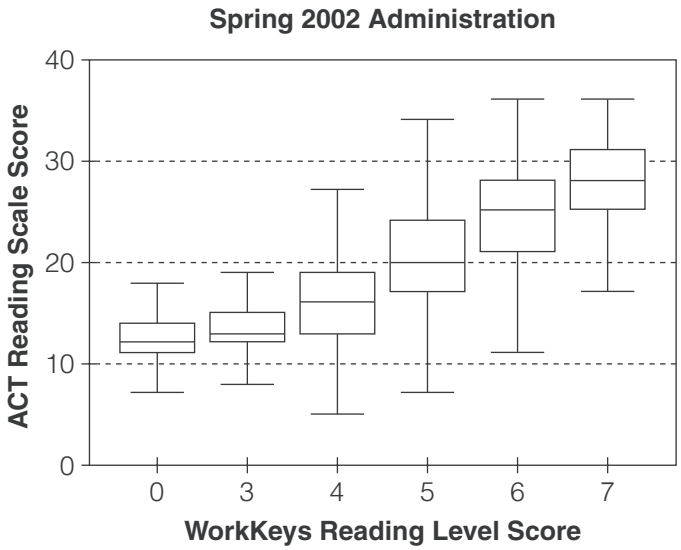
WorkKeys *Reading for Information*		ACT Reading Test						
		Below 16	16–19	20–23	24–27	28–32	33–36	Total
2002	Below 3	89.09	8.29	1.78	0.58	0.24	0.02	100
	3	80.01	16.46	3.03	0.40	0.09	0.01	100
	4	45.43	32.61	15.50	5.07	1.32	0.07	100
	5	14.61	27.47	28.70	18.66	9.12	1.45	100
	6	3.58	12.89	23.83	26.59	24.61	8.50	100
	7	0.57	3.45	10.47	21.26	35.71	28.55	100
	Total	30.85	24.08	19.22	13.22	9.34	3.31	100
2003	Below 3	89.74	7.26	1.94	0.75	0.24	0.07	100
	3	79.60	15.70	3.89	0.73	0.08	0.00	100
	4	49.88	28.66	14.91	5.13	1.38	0.04	100
	5	19.70	26.64	27.11	17.67	8.12	0.78	100
	6	5.61	13.57	23.67	27.91	24.41	4.82	100
	7	1.18	4.03	12.41	23.71	40.85	17.82	100
	Total	31.50	21.34	19.11	14.73	10.91	2.42	100

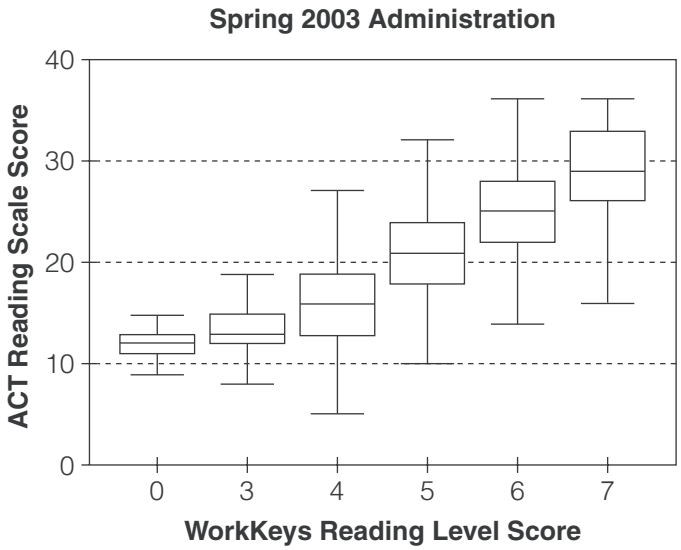
Figure 11 presents the conditional distributions of the Scale Scores for the ACT Reading Test given the Level Scores on WorkKeys *Reading for Information*. It shows the range (excluding extreme values), median, and quartile of the Scale Scores on ACT Reading for each *Reading for Information* Level Score. For example, for test takers who scored below Level 3 on *Reading for Information*, the actual observed range of Scale Scores on the ACT Reading Test in 2003 was 9 to 15, and the ACT Reading median Scale Score was 12.

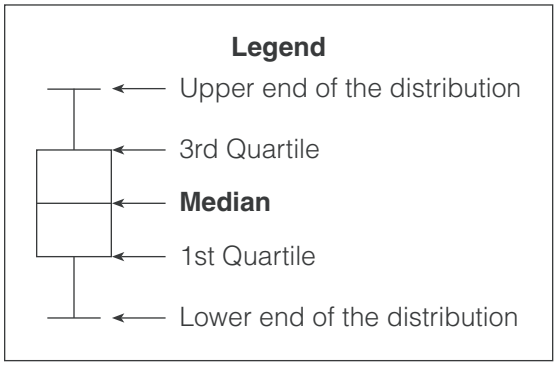
In summary, the results listed in Tables 19 and 20 and in Figure 11 show that there is a moderate relationship between WorkKeys *Reading for Information* and ACT Reading scores. In general, test takers who received higher Level Scores on WorkKeys *Reading for Information* received higher Scale Scores on the ACT Reading Test.

Figure 11

Boxplots of Scale Scores on ACT Reading at Each Level Score on WorkKeys *Reading for Information*







WorkKeys *Applied Mathematics* and the ACT Assessment

One would expect scores on WorkKeys *Applied Mathematics,* which focuses on the application of mathematical reasoning to work-related problems, to be related to other mathematics test scores. The relationship between WorkKeys *Applied Mathematics* and the ACT Mathematics Test provides construct evidence for test validity. The ACT Mathematics Test is part of the ACT Test, which is designed to measure the skills acquired during secondary education that are most important to success in postsecondary education. Thus, the ACT Mathematics Test measures the mathematical reasoning skills needed to solve practical problems in mathematics. The material the test covers emphasizes the major content areas that are prerequisite to successful performance in entry-level courses in college mathematics.

The results listed in Tables 21 and 22 and in Figure 12 show that WorkKeys *Applied Mathematics* scores are moderately correlated with ACT Mathematics scores. In general, test takers who received higher level scores on *Applied Mathematics* also received higher Scale Scores on the ACT Mathematics Test. These results imply that the abilities and/or skills measured by WorkKeys *Applied Mathematics* are similar to, but somewhat different from, those measured by ACT Mathematics.

Table 21
WorkKeys *Applied Mathematics* and ACT Mathematics Score Correlations

		ACT Mathematics Test	
		NC Score	**Scale Score**
	Range	1–60	1–36
2002	WorkKeys *Applied Mathematics*	0.81	0.75
2003	WorkKeys *Applied Mathematics*	0.78	0.71

Table 22 presents the conditional distributions of the ACT Mathematics Scale Score ranges for each Level Score on WorkKeys *Applied Mathematics*. Each row shows 100% of cases with the Level Score indicated. A value in a particular cell indicates the percentage of those cases that received an ACT Mathematics score within the range indicated for the column. For example, for both data sets, most of the test takers (about 89%) who scored below Level 3 on WorkKeys *Applied Mathematics* received Scale Scores below 16 on the ACT Mathematics Test.

Table 22

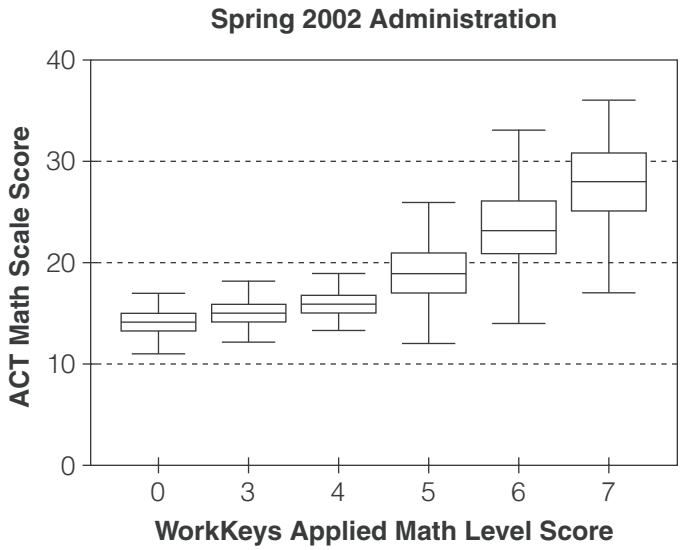
Percents of Test Takers by WorkKeys *Applied Mathematics* Level Scores and Ranges of ACT Mathematics Scale Scores

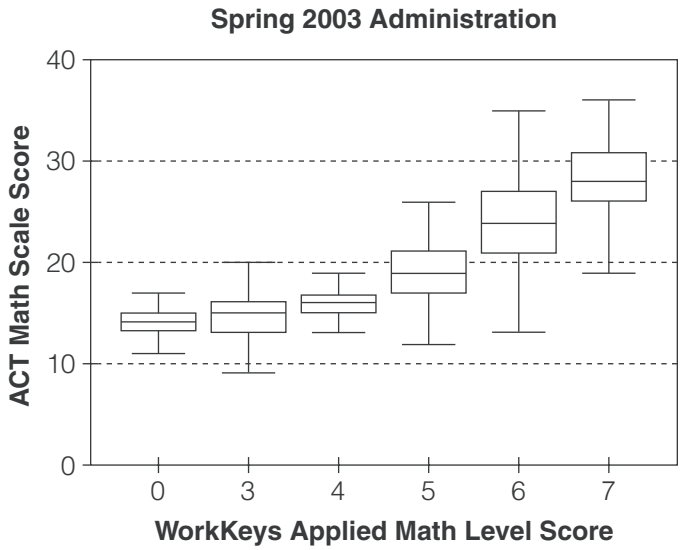
Level Scores on *Applied Math*		Scale Scores on the ACT Mathematics Test						
		Below 16	16–19	20–23	24–27	28–32	33–36	Total
2002	Below 3	89.75	9.77	0.36	0.10	0.01		100
	3	72.82	26.29	0.78	0.09	0.01	0.01	100
	4	37.37	54.75	7.11	0.72	0.04		100
	5	7.18	52.65	30.16	8.91	1.08	0.02	100
	6	0.62	16.14	34.47	33.11	14.36	1.30	100
	7	0.06	2.55	12.42	32.47	41.65	10.85	100
	Total	24.97	31.62	17.82	14.26	9.47	1.86	100
2003	Below 3	88.97	9.92	0.81	0.27	0.03		100
	3	74.01	25.21	0.71	0.04	0.03		100
	4	38.92	53.18	6.84	0.95	0.10		100
	5	9.06	51.18	27.82	10.59	1.30	0.04	100
	6	0.72	14.66	28.63	35.76	18.15	2.09	100
	7	0.06	1.49	9.67	32.39	44.30	12.09	100
	Total	25.46	32.73	17.61	14.92	7.99	1.29	100

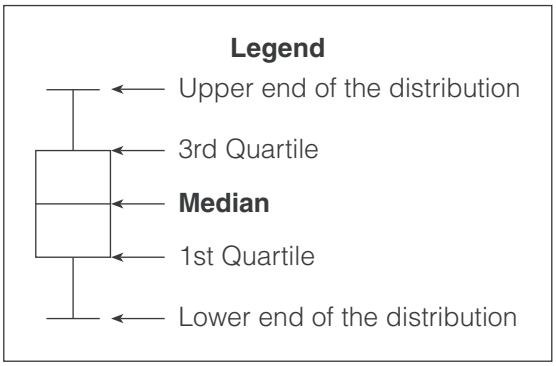
Figure 12 presents the conditional distributions of the Scale Scores for the ACT Mathematics Test given the Level Scores on WorkKeys *Applied Mathematics*. It shows the range (excluding extreme values), median, and quartile of the Scale Scores on ACT Mathematics for each *Applied Mathematics* Level Score. For example, for test takers who scored below Level 3 on *Applied Mathematics*, the actual observed range of Scale Scores on the ACT Mathematics Test was 11 to 17, and the ACT Mathematics median Scale Score was 14. In summary, the results listed in Tables 21 and 22 and in Figure 12 show that there is a moderate relationship between WorkKeys *Applied Mathematics* and ACT Mathematics scores. In general, test takers who received higher Level Scores on WorkKeys *Applied Mathematics* received higher Scale Scores on the ACT Mathematics Test.

Figure 12

Boxplots of Scale Scores on ACT Mathematics at Each Level Score on WorkKeys *Applied Mathematics*







Criterion-Related Evidence

The *Standards* (1999) pays particular attention to the area of employment skills testing, specifically stating that "the fundamental inference to be drawn from test scores in most applications of testing in employment settings is one of *prediction*: the test user wishes to make an inference from test results to some future job behavior or job outcome." The required behavior, which might be the satisfactory completion of an aspect of job performance, is commonly called a job *criterion*. For example, a required job behavior may be that an employee be able to read and interpret technical materials associated with the job. Criterion-related evidence might, therefore, show that a test taker with a certain score on a certain test can fairly be expected to read and interpret job-related technical materials that have a certain level of difficulty.

- *Criterion-related* evidence is collected by administering the tests to applicants or employees and comparing the results to supervisor ratings for the same people, thus comparing test results to observed behavior.
- In a *predictive* study, a test is administered to a group of job applicants but is not used for selection decisions. Some of the applicants are hired, and some are not. Performance data is later collected for those hired, and the test scores are compared to the performance data. This process makes it possible to gather information about the accuracy with which early test data can be used to estimate criterion scores that are obtained at a later time.
- In a *concurrent* study, a test is administered to job incumbents (hired applicants only) and the scores are compared to their current performance data. The test data and the performance data are collected in the same time period. There is no delay between the test administration and the collection of job performance ratings.

Performance is both multidimensional and dynamic (Borman, 1991; Austin & Villanova, 1992), so an employer's performance measures (e.g., performance ratings, absenteeism, tardiness) will, to some extent, include something other than true performance. These measures may therefore fail to include or accurately measure relevant aspects of true performance, and they may address additional issues that are not relevant to true performance (contaminating aspects). Thus, the results of a criterion-related validity study are likely to make an employment test look like a better or worse predictor than it really is. For example, performance ratings have been known to vary due to rater errors (Woehr & Huffcutt, 1994), and they frequently fail to include all of the dimensions that are important to evaluating performance (Schmidt, 1993). As a result, a study using performance ratings as the criterion will be limited in its ability to determine the true criterion-related validity of an employment test. Both measures are subject to random measurement error and this prevents both measures from having perfect reliability. The intercorrelation between two variables cannot exceed the reliability of either one. Therefore, the values of the reported correlations between the test scores and the job ratings can be attenuated due to the less-than-perfect reliability of the two measures.

Correlations between *Reading for Information* and Performance Ratings

As part of the process of validating WorkKeys tests for use in hiring decisions, the tests are administered to current (incumbent) employees. Then supervisors rate the same employees on their job performance, and the results are compared. The correlations between WorkKeys test scores and job performance ratings provide criterion-related evidence for the validity of using the specified WorkKeys test in relation to the specific job. The correlations reported here are based on incumbent employees; that is, only on applicants who were hired.

Table 23 presents an abbreviated selection of correlations between test scores and job performance ratings obtained from various organizations that studied the appropriateness of using WorkKeys for job applicant selection (the studies are numbered in the table for reference only). The jobs considered cover a wide spectrum. For example, they include machine operators, lab technicians, clerks, supervisors, and social workers. Both the sample size and the correlation vary from study to study. The correlations are positive and range from a low of .12 to a high of .86. Correlations of this type are typically considered to be very good if they are in the range of .2 to .3.

Table 23
Correlations between WorkKeys *Reading for Information* Scores and Job Performance Ratings

Study	Sample Size	Correlation
1	10	0.86
2	47	0.58
3	31	0.51
4	19	0.47
5	30	0.43
6	26	0.39
7	56	0.38
8	27	0.34
9	142	0.33
10	21	0.26
11	36	0.17
12	103	0.16
13	120	0.16
14	173	0.14
15	314	0.12

Correlations between *Applied Mathematics* and Performance Ratings

Table 24 presents an abbreviated selection of correlations between test scores and job performance ratings obtained from various organizations studying the appropriateness of using WorkKeys for job applicant selection. The jobs considered cover a wide spectrum. They include machine operators, lab technicians, clerks, supervisors, and social workers. Both the sample size and the correlation vary from study to study. All of the correlations are positive and provide positive criterion-related validity evidence for using WorkKeys as part of the job applicant selection process. According to the U.S. Department of Labor Employment and Training Administration, correlations above .35 are "very beneficial" and those from .21 to .35 are "likely to be useful" (U.S. Department of Labor and Training Administration, 2000).

Table 24
Correlations between WorkKeys *Applied Mathematics* Level Scores and Job Performance Ratings

Study	Sample Size	Correlation
1	142	0.41
2	27	0.41
3	24	0.41
4	141	0.41
5	56	0.34
6	120	0.23

Correlations between *Locating Information* and Job Performance Ratings

Table 25 presents an abbreviated selection of correlations between test scores and job performance ratings obtained from various organizations studying the appropriateness of using WorkKeys for job applicant selection. The jobs considered cover a wide spectrum. They include machine operators, lab technicians, clerks, supervisors, and social workers. Both the sample size and the correlation vary from study to study. All of the correlations are positive and provide positive criterion validity evidence for using WorkKeys as part of the job applicant selection process.

Table 25
Correlations between WorkKeys *Locating Information* Level Scores and Job Performance Ratings

Study	Sample Size	Correlation	Study	Sample Size	Correlation	Study	Sample Size	Correlation
1	13	0.42	7	56	0.26	13	26	0.18
2	47	0.41	8	19	0.23	14	173	0.17
3	42	0.32	9	39	0.22	15	30	0.15
4	22	0.31	10	27	0.21	16	314	0.14
5	126	0.30	11	36	0.20			
6	70	0.29	12	120	0.19			

Classification Consistency

Another way to measure the criterion validity of WorkKeys tests is to examine the percentage of workers correctly classified by the tests (see Table 26). Incumbent employees are classified into groups of successful and less-successful employees based on their supervisor's rating of their job performance. After taking the WorkKeys tests, the employees can also be classified according to their scores on the tests. If they achieved the minimum acceptable scores, they are classified as successful, otherwise as not successful. Comparing the employees' job performance classification with their WorkKeys test classification yields a measure of classification consistency for the WorkKeys tests. Correctly classified employees are those who were classified the same way by both measures. That is, the total number of correctly classified employees is the number classified as successful by both measures plus the number classified as unsuccessful by both measures. Employees who are not given the same classification by both measures are misclassified.

Classification consistency is typically underestimated due to restriction of range and the less-than-perfect reliability of the two measures. It can also be affected if the cutoff score is set too high or low. For hiring decisions, a passing score generally represents the minimum acceptable skill level. In other situations, however, the passing score might be set at a skill level that is more desirable and does not represent minimum requirements. This has the effect of reducing the percent of correct classifications when incumbents are tested. The following tables should be viewed with these caveats in mind.

Reading for Information

Table 26 presents an abbreviated selection of the percentages of correct classifications resulting from *Reading for Information* test scores and job performance ratings obtained from various businesses that studied the appropriateness of using WorkKeys tests for job applicant selection.

The first column in Table 26 shows the number (N) of participants in each study. The passing score for each study is shown in the second column, and the third column shows the percent of employees (of N) whose performance was rated as "successful" and whose scores were at or above the passing score. For example, the first study included a total of 33 test takers. Of these, 79% were rated as successful and scored at or above Level 3, which was the cutoff score for the specified job.

Table 26
Job Classification Consistency with *Reading for Information*

N	WorkKeys Passing Score	Percent Correctly Classified
33	Level 3	79
120	Level 6	75
103	Level 5	73
56	Level 3	71

Applied Mathematics

Table 27 presents an abbreviated selection of the percentages of correct classifications between test scores and job performance ratings obtained from various businesses studying the appropriateness of using WorkKeys for job applicant selection. The sample size, denoted by N, is given in the first column and the passing score on the WorkKeys test is given in the second column. The third column shows the percentage of employees whose performance was "successful" and whose scores were at or above the passing score.

Table 27
Job Classification Consistency with *Applied Mathematics*

N	WorkKeys Passing Score	Percent Correctly Classified
33	Level 3	79
120	Level 4	90
56	Level 4	57

Locating Information

Table 28 presents an abbreviated selection of the percentages of correct classifications between test scores and job performance ratings obtained from various businesses studying the appropriateness of using WorkKeys for job applicant selection. The sample size, denoted by N, is given in the first column and the passing score on the WorkKeys test is given in the second column. The third column shows the percentage of employees placed at the indicated level on both measures.

Table 28
Job Classification Consistency with *Locating Information*

N	WorkKeys Passing Score	Percent Correctly Classified
20	Level 4	100
39	Level 4	79
126	Level 4	88
56	Level 5	30
120	Level 5	37

Content-Related Evidence

The *Uniform Guidelines* (1978) indicates that employers using a content validation strategy should focus on observable work behaviors as the aspect of job performance to which they link the content of the test. The test may measure a type of knowledge, skill, or ability needed to perform these observable work behaviors. However, the *Uniform Guidelines* states that a content validation approach is not sufficient for justifying the use of tests that measure mental processes (e.g., common sense, personality) that are not directly observable or discernible through observable work behaviors. The *Standards* states that content-related validity evidence used for selecting, promoting, or classifying employees should be based on a job analysis that defines the work performed. The *Uniform Guidelines* further specifies that the job analysis should yield information regarding the critical work behaviors or tasks that the job comprises.

For the WorkKeys assessments, content-related validity evidence can be established in employment settings by linking test scores to the set of job behaviors or job outcomes of interest. The WorkKeys job profiling procedure enables trained job analysts or profilers to conduct a job analysis to document content-related validity evidence for each WorkKeys assessment. Test scores can also be linked to job behaviors using SkillMap®. ACT developed SkillMap for users who have difficulty accommodating the focus group meetings used in job profiling. In this way, content-related validity evidence is documented, and the skill levels identified as required can then be used as criteria—cutoff scores—on the specified WorkKeys assessments.

The job profiling procedure and SkillMap are both designed to meet the standards for content validation established in the *Uniform Guidelines*. Both of them are used to:

- define the critical job tasks,
- determine which WorkKeys skills are relevant to performing the tasks, and
- identify the level of skill required for performing them.

Figure 13 shows how these WorkKeys job analysis methods meet federal and professional standards for establishing validity evidence in high-stakes situations. Figure 13 mentions ACT job profilers, subject matter experts (SMEs), and job experts. Job profilers are job analysts trained by ACT to conduct the WorkKeys job profiling procedure. During the profiling procedure, the job profilers work in focus groups with SMEs. After meeting with the focus groups, job profilers prepare a report that includes the job profile data. When SMEs are mentioned in Figure 13, this means that the job profiling procedure is being described.

Like SMEs, job experts are people who are familiar with the specified job. They participate in the SkillMap process without meeting as a group. When job experts are mentioned, this means that SkillMap is being described. SkillMap is a Web-based program. When the job experts have completed their SkillMap tasks, SkillMap can generate a job inventory. A WorkKeys job profile and a SkillMap job inventory are two different things that are produced through two different procedures, but both can be used to establish skill level requirements for a job and contribute to the overall validity evidence for the test.

Figure 13

Comparison of the *Uniform Guidelines* Requirements and Two ACT WorkKeys Job Analysis Procedures for Content Validation

***Uniform Guidelines* Requires**	**WorkKeys Procedures**
A job analysis that generates descriptions of job behaviors, descriptions of tasks, and measures of their criticality	SMEs (participating in the job profiling procedure) or job experts (using SkillMap) establish a list that describes behaviors and tasks; then they rate each task for *Importance* and *Relative Time Spent* in order to yield a *Criticality* rating for each task.
Demonstration that the test is related to described job behaviors and tasks	ACT job profilers report the percentage of tasks that require the skill. SkillMap job inventory software lists the tasks linked to a skill and shows the number of job experts who linked each task to the skill.
Definition of skills in terms of observable work outcomes	Each WorkKeys skill and skill level is defined with specific criteria and is illustrated with multiple workplace examples. SMEs or job experts link these definitions to job behaviors and tasks.
Explanation of how the skills are used to perform the tasks or behaviors	SMEs or job experts identify tasks that require the skill under review. SMEs link specific tasks to a skill level and say how the level is used for the tasks. Job experts assign tasks to skills and skill levels.
That no decisions be made based on knowledge, skills, and abilities that can be learned quickly on the job or in training	SMEs identify the skill level required for job entry. New hires should enter the job with this level, not learn it on the job. Job experts identify tasks performed at job entry and link them to skill levels. An algorithm compiles the results.
That applicants be assessed on skills for higher-level jobs only if new hires may advance quickly	SMEs identify the skill level required for performing the job on the first day. They may, in addition, set a higher skill level for performing the job effectively after training.
That the rationale for setting cutoff scores must be provided	SMEs identify cutoff skill levels by describing job tasks and linking skill level descriptions and examples to them. SkillMap uses an algorithm to set cutoff scores based on task criticality and the SMEs' assignment of tasks to skill levels.
That the cutoff scores used are to be consistent with normal expectations of workers	SMEs identify the cutoff skill level based on the normal requirements of the job, not on unusual job situations, desired capabilities, or their beliefs regarding their own skill levels.
That results supporting pass/fail decisions only, must not be used to rank test takers	WorkKeys scores show that test takers either have the required skill levels or do not have them. It is not appropriate to rank applicants based on their WorkKeys scores.
That documentation regarding validation efforts is to be maintained	Job profilers present a full report documenting content-related validity evidence, and they retain all related worksheets and computer records. SkillMap generates content-related validity documentation and a thorough record of the entire process. Users may download SkillMap data.

Adverse Impact

The *Uniform Guidelines* indicates that the use of employment tests does not violate federal law unless it is found that use of the test results in "adverse impact on employment opportunities of any race, sex, or ethnic group" (p. 203) and the user cannot demonstrate its business necessity (see Title VII of the *Civil Rights Act* of 1964). The *Uniform Guidelines* explains the four-fifths or 80% rule of thumb for determining adverse impact. To estimate whether adverse impact exists, an employer compares the ratio of minority group applicants hired to the ratio of majority group applicants hired. Adverse impact is considered to be present when the ratio for the minority group is less than four-fifths, or 80%, of the ratio for the majority group. That is,

- *if 100% of the applicants from a majority group are hired, then adverse impact is indicated if less than 80% of the applicants from a minority group are hired.*

However, this rule of thumb is meant as a practical tool for estimating the presence of adverse impact and is not a definitive test.

The *Uniform Guidelines* also indicates that when the use of an employment assessment results in adverse impact, an employer may continue to use the assessment after

1. demonstrating its business necessity by validating its use, and
2. demonstrating that the use of similar tests will not result in less adverse impact.

Gender and Race/Ethnicity Analyses

Reading for Information

Table 29 presents data comparing Level Scores for male and female test takers who took *Reading for Information.* There were statistically significant differences between males and females, but because only integer Level Scores are reported, both groups effectively scored at the same level. In addition, applying a criterion of a difference in Level Score of 0.5 or more, no practical score differences were detected between males and females. Because there is a potential for adverse impact with any cognitive ability test, employers should make sure that a well documented job analysis links the job to the skills and the skills to the assessment tool. The cutoff score should be set at a level that is clearly appropriate and the reasons for using that score should also be well documented.

The fact that statistically significant differences in cognitive ability test performance are typical between a majority and a minority group (for example, 1 SD difference between Caucasians and African Americans) has been thoroughly researched and documented (Ryan, 2001). Performance on the WorkKeys assessments is consistent with these findings. A review of Table 29 also shows that statistically significant differences in test-taker scores for *Reading for Information* by race/ethnicity were detected.

Table 29

Descriptive Statistics of *Reading for Information* Mean Level Scores by Gender and Race/Ethnicity

		N	Mean	Standard Deviation
Gender	Female	627,236	4.60	1.203
	Male	632,084	4.38	1.369
Race/ Ethnicity	African American/Black, Non-Hispanic	249,720	4.06	1.253
	Asian American or Pacific Islander	27,488	4.53	1.338
	Caucasian/White, Non-Hispanic	719,758	4.74	1.225
	Hispanic/Latino	43,248	3.89	1.469

With both the gender analysis and the race/ethnicity analysis, it is important to look at practical differences. A difference in mean Level Score of 0.5 or more among the four race/ethnic groups was considered practically significant. A one-way ANOVA was used to compare each of the groups. The ANOVA indicated a significant difference among the four groups.

Using the performance level difference of 0.5 or more, results of a Bonferroni Post Hoc test determined that there were statistically significant and practical mean differences between Caucasians and African Americans, between Caucasians and Hispanic/Latinos, and between Asian Americans and Hispanic/Latinos. While these findings are consistent with existing research, an employer's use of any assessment for employment decisions should be clearly linked to the critical tasks required for the job. The task and WorkKeys skill requirements for a job can be established through a job analysis and validity study using, for example, WorkKeys job profiling or SkillMap. The results of such a study will establish which assessments are appropriate for employment decisions.

Applied Mathematics

Table 30 presents data comparing Level Scores for male and female test takers who took *Applied Mathematics.* There were statistically significant differences between males and females, but because only integer Level Scores are reported, both groups effectively scored at the same level. In addition, applying a criterion of a difference in Level Score of 0.5 or more, no practical score differences were detected between males and females.

Table 30 also shows that statistically significant differences in test-taker scores by race/ethnicity were detected for *Applied Mathematics.*

Table 30
Descriptive Statistics of *Applied Mathematics* Mean Level Scores by Gender and Race/Ethnicity

		N	Mean	Standard Deviation
Gender	Female	635,325	4.22	1.422
	Male	655,645	4.40	1.509
Race/ Ethnicity	African American/Black, Non-Hispanic	255,121	3.42	1.354
	Asian American or Pacific Islander	26,636	4.83	1.526
	Caucasian/White, Non-Hispanic	750,409	4.72	1.341
	Hispanic/Latino	42,619	3.65	1.517

A difference in mean Level Score of 0.5 or more among the four race/ethnic groups was considered practically significant. A one-way ANOVA was used to compare each of the groups. The ANOVA indicated a significant difference among the four groups.

Using the performance level difference of 0.5 or more, results of a Bonferroni Post Hoc test determined that for each of the WorkKeys assessments, there were statistically significant and practical mean differences between Caucasians and African Americans, between Caucasians and Hispanic/Latinos, and between Asian Americans and Hispanic/Latinos. While these findings are consistent with existing research, an employer's use of any assessment for employment decisions should be clearly linked to the critical tasks required for the job. The task and WorkKeys skill requirements for a job can be established through a job analysis and validity study using, for example, job profiling or SkillMap. The results of such a study will establish which assessments are appropriate for employment decisions.

Locating Information

Table 31 presents data comparing Level Scores for male and female test takers who took *Locating Information.* There were no statistically significant differences between males and females; both groups effectively scored at the same level. In addition, applying a criterion of a difference in Level Score of 0.5 or more, no practical score differences were detected between males and females.

Table 31 shows that statistically significant differences in test-taker scores by race/ethnicity were detected for *Locating Information.*

Table 31
Descriptive Statistics of *Locating Information* Mean Level Scores by Gender and Race/Ethnicity

		N	Mean	Standard Deviation
Gender	Female	316,254	3.49	1.113
	Male	382,298	3.49	1.185
Race/ Ethnicity	African American/Black, Non-Hispanic	140,294	3.04	1.203
	Asian American or Pacific Islander	8,469	3.19	1.285
	Caucasian/White, Non-Hispanic	375,522	3.73	1.040
	Hispanic/Latino	30,659	3.08	1.344

Using the performance level difference of 0.5 or more, results of a Bonferroni Post Hoc test determined that there were statistically significant and practical mean differences between Caucasians and African Americans, between Caucasians and Hispanic/Latinos, and between Caucasians and Asian Americans. While these findings are consistent with existing research, an employer's use of any assessment for employment decisions should be clearly linked to the critical tasks required for the job. The task and WorkKeys skill requirements for a job can be established through a job analysis and validity study using, for example, WorkKeys job profiling or SkillMap. The results of such a study will establish which assessments are appropriate for employment decisions.

WorkKeys Job Analysis Options

WorkKeys job analysis tools aid in the identification of skills and skill levels that current and prospective employees need for success on the job. ACT WorkKeys offers three job analysis options for setting skill level standards on the WorkKeys assessments: WorkKeys job profiling, SkillMap job inventory, and WorkKeys Estimator (used only for estimating skill levels, not establishing them with evidence).

Job Profiling

The WorkKeys job profiling procedure is a method of job analysis conducted by analysts who have been trained and authorized by ACT industrial/organizational psychologists.

Job profilers. ACT offers Job Profiling Training to qualified individuals. The training consists of several weeks of distance learning activities culminating in an on-site workshop. The analysts are trained to develop task lists and conduct job profiling sessions at job sites. In the sessions, SMEs provide information and explanations about how the job tasks require specified skills and skill levels.

Subject matter experts. The SMEs are individuals who are familiar with the job being studied. They typically include job incumbents and may include their supervisors or other employees who are familiar with the job.

Skill levels. The outcome of the procedure is a set of recommended and prioritized standards describing the skill levels required for job entry and effective performance. These skill levels correspond directly to scores on the WorkKeys tests.

Report. After the profile sessions for a job have been completed, the profiler prepares a written report that indicates which skills and skill levels are relevant to that job and lists them according to their criticality to the job. The report includes a task list and it links the tasks to the skills and skill levels needed to perform those tasks. Thus, the report provides documentation of content validity evidence and facilitates the use of the WorkKeys assessments for personnel selection and/or promotion purposes.

SkillMap Job Inventory

SkillMap is a Web-delivered job analysis system that links the tasks of the specified job to the WorkKeys skills and skill levels. SkillMap does not require the participation of an ACT-authorized job profiler.

Administrator. The SkillMap procedure is facilitated by a local administrator who can coordinate SkillMap activities by using the instructions that are built into SkillMap. Special training is not required, but ACT-authorized job profilers often have additional expertise to offer.

Job experts. The administrator contacts job experts and informs them about the activities they are asked to complete. SkillMap guides the job experts through the process of identifying job tasks and the WorkKeys skills and skill levels needed for completing those tasks.

Job inventory report. After the job tasks and skill levels have been entered, the software produces the SkillMap Job Inventory Report. The job inventory lists the required WorkKeys skills and skill levels and indicates how critical they are to the job. The skills and skill levels correspond to the WorkKeys assessments and cutoff scores. In addition, the report provides content validity evidence that (a) documents the appropriate use of the WorkKeys assessments for personnel selection and training or promotion purposes, and that (b) complies with the report requirements outlined in the *Uniform Guidelines on Employee Selection Procedures* (1978).

WorkKeys Estimator

This step-by-step process is designed to provide users with a method of documenting their decisions concerning the use of WorkKeys assessments. Companies may use WorkKeys Estimator to assist with low-stakes uses of the WorkKeys assessments such as enhancing recruiting efforts and developing training goals.

Coordinator. WorkKeys Estimator is facilitated by a coordinator who does not need to be trained by ACT and can coordinate the procedure using the instructions that are built into WorkKeys Estimator. The coordinator communicates with management and with the job experts, and collects data, manages the flow of information, but does not make decisions.

Job experts. Job experts are individuals who are knowledgeable about the job and how it is performed. The job experts work with the coordinator to independently review WorkKeys skill and skill level definitions. Based on their knowledge of the job tasks, the job experts document their estimate of the skill levels needed for completing them.

Management. Responsible decision makers review the job experts' estimates along with additional information from the WorkKeys occupational profiles. Based on all of the information collected by the coordinator and based on the recommendations included in the WorkKeys Estimator documentation, management decides which skill level estimates to use for a job.

While this process does include a review of job information and skill information by job experts, it generates skill level estimates only. It does not create task lists that link skill levels to the tasks of the job. If these are needed for high-stakes decisions, employers should use ACT WorkKeys job profiling or SkillMap.

References

American Management Association. (2001). *AMA survey on workplace testing: Basic skills, job skills, psychological measurement.* New York: Author.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77*(6), 836–874.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 271–326). Palo Alto, CA: Consulting Psychologists Press.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Civil Rights Act. (1964). Document number PL88-352. Washington, DC: U.S. Government Printing Office.

Cronbach, L. J., Gleser, G. C., Nanda, H. I., & Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability of scores and profiles.* New York: Wiley.

Deloitte Development. (2005). *2005 Skills Gap Report—A Survey of the American Manufacturing Workforce.* New York: Author.

Freeman, M. F., & Tukey, J. W. (1950). Transformation related to the angular and square root. *The Annals of Mathematical Statistics, 21*, 607–611.

Gao, X., Chen, H., & Harris, D. J. (2005). *Consistency of Equating Functions Across Different Equating Designs, Methods and Samples: An Empirical Investigation*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Gao, X., Harris, D. J., Yi, Q., & Lei, M. (2003). *Examining consistency of item parameters estimated in pretest and operational test administrations.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. A. Lasarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and Prediction* (pp. 60–90). Princeton: Princeton University.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29–36.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for Scale Scores. *Journal of Educational Measurement, 29*, 285–307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for Scale Scores using IRT. *Journal of Educational Measurement, 33*, 129–140.

Lee, W., Brennan, R. L., & Hanson, B. A. (2000). *Procedures for computing classification consistency and accuracy indices with multiple categories.* ACT Research Report Series. Iowa City, IA: ACT.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 221–262). New York: American Council on Education, and Macmillan.

Ryan, A. M. (2001). Explaining the black-white test score gap: The role of test performance. *Human Performance, 14*(1), 45–75.

Schmidt, F. L. (1993). Personnel psychology at the cutting edge. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 497–516). San Francisco: Jossey-Bass.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1997). A study of modified-Guttman and scale scores using IRT. *Journal of Educational Measurement, 33*, 129–140.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*, 347–362.

Standards for Educational and Psychological Testing (1999). Washington, DC: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 267–290). Baltimore: The Johns Hopkins University Press.

U.S. Department of Education National Center for Education Statistics (2003). *The condition of education 2003* (NCES 2003-067). Washington, DC: Author.

U.S. Department of Labor Employment and Training Administration (2000). *Testing and assessment: An employers guide to good practice*. Washington, DC: Author.

U.S. Department of Labor—Bureau of Labor Statistics (2002, August 27). *Number of jobs held, labor market activity, and earnings growth among younger baby boomers: Results from more than two decades of a longitudinal survey* (USDL 02-497). Washington, DC: Author.

U.S. Department of Labor—Bureau of Labor Statistics (2003, June 25). *College enrollment and work activity of 2002 high school graduates* (USDL 03-330). Washington, DC: Author.

Uniform guidelines on employee selection procedures, (1978). U.S. Equal Employment Opportunity Commission. *Federal Register, 43*, 38290–38315.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational & Organizational Psychology, 67*(3), 189–205.

Zimowski, M. F., Muraki, E., Mislevy, R. J., Bock, R. D. (1996). *BILOG-MG: Multiple-group item analysis and test scoring.* Chicago: Scientific Software International.