



Assessment Service Bulletin Number I

WJ IV™ Tests of Achievement Alternate-Forms Equivalence

Erica M. LaForte, PhD

Kevin S. McGrew, PhD

This bulletin is an executive summary and discussion of all extant alternate-forms equivalence data gathered for the WJ IV Tests of Achievement (Schrack, Mather, & McGrew, 2014). The WJ IV Technical Manual (McGrew, LaForte, & Schrank, 2014) contains information about how the three forms of each test were constructed to be equivalent in content coverage, difficulty, and measurement error. A summary of these procedures is included in Part A of this bulletin. Part B includes the descriptions and results of two recently completed alternate-forms studies. These studies contribute to the accumulation of evidence regarding the equivalence of the WJ IV ACH Standard tests. Additionally, these studies present an example of methodology appropriate for evaluating empirical evidence for alternate-forms equivalence in the context of individually administered assessments.



Houghton
Mifflin
Harcourt

Copyright © 2014 by The Riverside Publishing Company. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording or by any information storage or retrieval system, without the prior written permission of Riverside, unless such copying is expressly permitted by federal copyright law. Address all inquiries to Permissions, Riverside, 3800 Golf Road, Suite 200, Rolling Meadows, Illinois 60008.

Printed in the United States of America.

Woodcock-Johnson is a registered trademark of Houghton Mifflin Harcourt Publishing Company. WJ IV and the Woodcock-Johnson IV logo are trademarks of Houghton Mifflin Harcourt Publishing Company.

Reference Citation

- To cite this document, use:

LaForte, E. M., & McGrew, K. S. (2014). *WJ IV Tests of Achievement Alternate-Forms Equivalence* (Woodcock-Johnson IV Assessment Service Bulletin No. 1). Rolling Meadows, IL: Riverside.

For technical information, please visit www.riversidepublishing.com or call Riverside Customer Service at 800.323.9540.

1 2 3 4 5 6 7 8 9 10-XXX-18 17 16 15 14



WJ IV™ Tests of Achievement Alternate-Forms Equivalence

The *Woodcock-Johnson® IV Tests of Achievement* (WJ IV ACH) (Schrang, Mather, & McGrew, 2014) includes three forms of the eleven tests in the Standard Battery (Forms A, B, and C). In some educational settings, alternate forms are used for different purposes or are used by different examiners. In cases where an independent educational evaluator is contracted by a student's parents or by a school in a disputed evaluation, use of an alternate form is preferred. In addition, the availability of alternate forms makes the WJ IV ACH tests useful for repeated testing of an individual over time, such as pretest-posttest research designs or in educational programs requiring subsequent test administrations to measure progress or change over time. By using a different, but equivalent, form for each administration, professionals can reduce dependence on any single form of the test and minimize potential overexposure to the test items (practice effect). The three forms of the WJ IV ACH tests were carefully constructed to be as equivalent as possible.

There are several fundamental criteria to consider when assessing the equivalence of alternate forms:

1. *Equivalence of content coverage.* Does each form of the test include approximately the same proportion of items representing each curricular strand?
2. *Equivalence of form difficulty.* Is the average difficulty of the items on each form approximately equal?
3. *Equivalence of measurement precision.* Do the forms exhibit approximately equal standard errors of measurement (*SEM*) across the entire range of ability?
4. *Equivalence of rank-ordering of examinee ability.* Do examinees who perform well on one form of the test also perform well on the other forms?

Examiners who wish to compare an examinee's scores on alternate forms should evaluate whether existing evidence supports each type of equivalence listed above. In practice it is nearly impossible for two forms of a test to be exactly equivalent in all characteristics. Thus, it is important that test users understand how different types of equivalence evidence may impact score interpretations and score comparisons among forms. For example, it is possible that examinee scores from two forms of a test might be highly correlated, indicating that the forms similarly rank order individuals within that sample of examinees even though one form of the test might yield significantly lower examinee scores than the other form. This pattern of evidence might suggest that the two forms are equivalent in content coverage but that one form contains items that are systematically more difficult for most examinees. In another example, two forms of a test might yield similar examinee mean scores, providing evidence for equivalence of form difficulty, but one form might have a much larger standard error of measurement across the ability range, indicating that it is less precise in locating an examinee's ability on the underlying scale.

Standard 4.10 in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) states that, “A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably” (p. 57). Evidence to support the equivalence of alternate test forms should begin with documentation from the test developer about the methods used for item equating and form construction and should continue to be gathered throughout the life cycle of the test in the form of post-publication studies of equivalence. Ideally, operational equivalence studies should be conducted across several different samples of varying age, ability, and demographic characteristics, and in multiple settings, so that test users can evaluate the empirical evidence for form equivalence in a meta-analytic fashion.

The following sections document the alternate-forms equivalence evidence gathered to date for the WJ IV ACH Standard Battery tests. A summary of the form construction processes employed during the development of the WJ IV is included in Part A. Part B includes the descriptions and results of two recently completed alternate-forms studies.

Part A: Documentation of Processes for WJ IV ACH Alternate-Forms Construction

The *WJ IV Technical Manual* (McGrew, LaForte, & Schrank, 2014) contains detailed information about the item pool calibration and equating procedures used during the WJ IV norming study. These processes resulted in large item pools, ranging from 100 to 600 items calibrated to common *W* scales using data from between 2,500 and 25,000 individuals collected during the WJ IV norming study and the development of prior editions of the Woodcock-Johnson tests.

WJ IV ACH Standard Nonspeeded Tests

Test 1: Letter-Word Identification, Test 2: Applied Problems, Test 3: Spelling, Test 4: Passage Comprehension, Test 5: Calculation, Test 6: Writing Samples, Test 7: Word Attack, and Test 8: Oral Reading are nonspeeded tests in the WJ IV ACH. Using Rasch item banking and equating methodology, items in each of these pools were calibrated and placed onto a common *W*-difficulty scale. Items were then selected from the item pools to appear on the published Form A, Form B, and Form C of the WJ IV ACH tests using the following guidelines.

Guideline 1: Equivalence of Content Coverage

Items were selected so that each form contained equal representation of the intended breadth of content for that test. Content-area curriculum experts provided consultation on the comparability of the item types for key criteria in Test 1: Letter-Word Identification, Test 2: Applied Problems, Test 3: Spelling, Test 5: Calculation, and Test 7: Word Attack. For example, a content-area expert reviewed all three forms of Test 3: Spelling to ensure that the forms contained equivalence in the numbers of items that included different types of phonetic components such as vowel digraphs, *r*-controlled blends, silent letters, and doubling consonants. In a similar manner, a mathematics curriculum expert verified that each form of Test 5: Calculation contained approximately equal numbers of items measuring basic math concepts such as addition, subtraction,

multiplication, and division as well as more advanced concepts such as fractions, algebra, geometry, trigonometry, and factorials. Where necessary, adjustments were made to item selections to ensure content equivalence across forms.

Table 1 contains an example of the application of guideline 1—equivalence of item content—in the development of Test 5: Calculation. This table shows the distribution of item types for each form of the test. Each form contains equal, or nearly equal, numbers of items representing the entire breadth of the item pool for this test. Adherence to the guideline of equivalence of item content ensures that no examinee encounters more or fewer items from a certain curricular strand than examinees who take other forms of the test.

Table 1.
*Frequency of Item Types
Across Three Forms
of WJ IV ACH Test 5:
Calculation*

Item Type	Number of Items per Form		
	Form A	Form B	Form C
Basic Addition	12	10	10
Basic Subtraction	6	8	8
Basic Multiplication	4	4	4
Basic Division	2	3	2
Advanced Addition	3	3	4
Advanced Subtraction	1	1	2
Advanced Multiplication	3	3	4
Advanced Division	2	2	2
Advanced Math	1	1	1
Algebra	5	4	4
Derivatives	2	2	2
Factorials	1	1	1
Fractions	4	4	4
Geometry	1	1	1
Integrals	2	3	3
Logarithms	1	1	1
Matrices & Determinants	1	1	1
Percentages	2	1	1
Powers & Roots	2	2	1
Trigonometry	2	2	1

Guideline 2: Equivalence of Form Difficulty

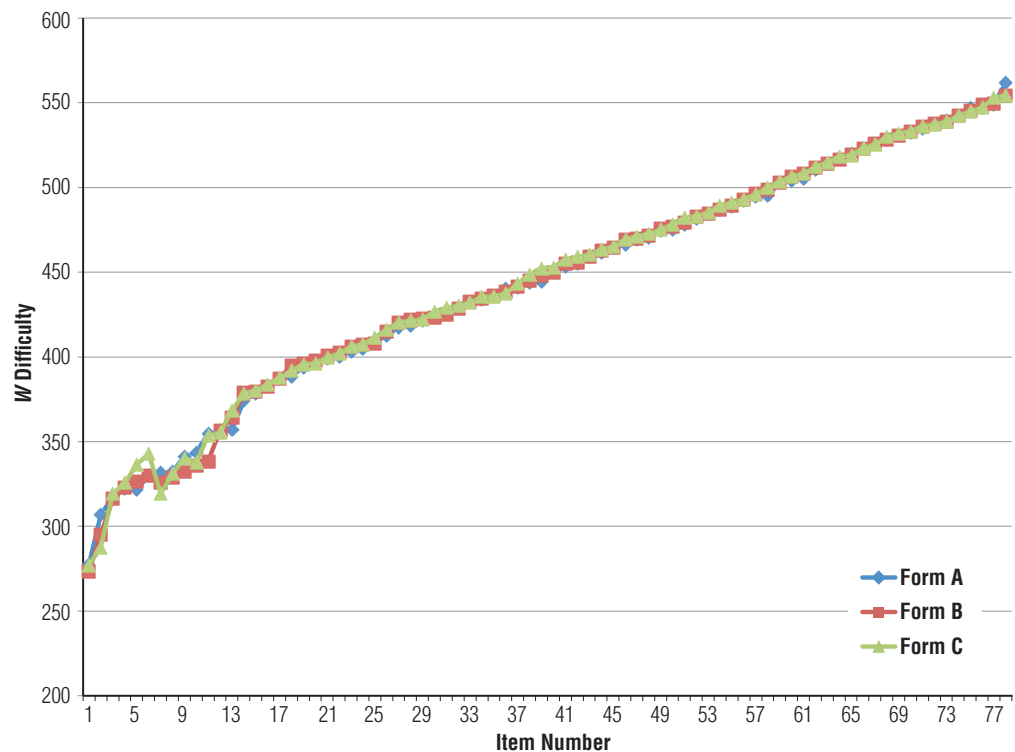
Items were selected for each form so that the item difficulty gradient, or number of items per 10 W points of ability, was approximately equal for each form. This guideline ensures that there are no obvious gaps in item difficulty on any of the test forms.

Figures 1 through 8 present evidence that this guideline was met in the construction of Tests 1 through 8. These figures plot the W difficulties of the items, in serial order, for Forms A, B, and C of the tests. In almost all forms of Tests 1 through 7, the number of items per 10 W points is equivalent across the entire range of W ability covered by the test. Additionally, most corresponding item difficulties are equivalent within 2 W points across forms. There are a few exceptions, however, in cases where item pool depth or content coverage guidelines precluded an exact match between corresponding

item difficulties.¹ For instance, at the raw score range 19 through 23 in Test 2: Applied Problems (see Figure 2), Form A contains items with difficulties approximately 5 to 10 W points lower than the corresponding items on Forms B and C. Also, in the raw score range 12 through 22 in Test 4: Passage Comprehension (see Figure 4), Form C contains items with difficulties approximately 5 to 10 W points lower than the corresponding items in Forms A and B. In these cases, although great care was taken to ensure equivalence in corresponding item difficulty, small tradeoffs in item difficulty difference were made to maintain the equivalence in breadth of item types across three forms.

Test 8: Oral Reading is administered in item sets; in each form of the test, sets of items together form meaningful “stories.” The items in this test are calibrated onto a common scale in the same manner as the items in the other WJ IV tests are calibrated. However, because of the story-based administration and because this test is new in the WJ IV, the item pool underlying the Oral Reading test contains a more limited number of items than the item pools for the other WJ IV ACH Standard tests. For this reason, the items on the three forms of Oral Reading are expected to, and do, vary in difficulty more than the items on Tests 1 through 7 do because Tests 1 through 7 are constructed from much larger pools of standalone items. The slight variation in item difficulty between forms is evident in Figure 8.

Figure 1.
*Plot of item W difficulties
across three forms of
WJ IV ACH Test 1: Letter-
Word Identification.*



¹ In instances where item W difficulties are not exactly the same across the three forms, the scoring tables used to convert raw scores to W-ability scores adjust for these differences. This is discussed in a later section and visually depicted in Figures 9 through 16.

Figure 2.
*Plot of item W difficulties
 across three forms
 of WJ IV ACH Test 2:
 Applied Problems.*

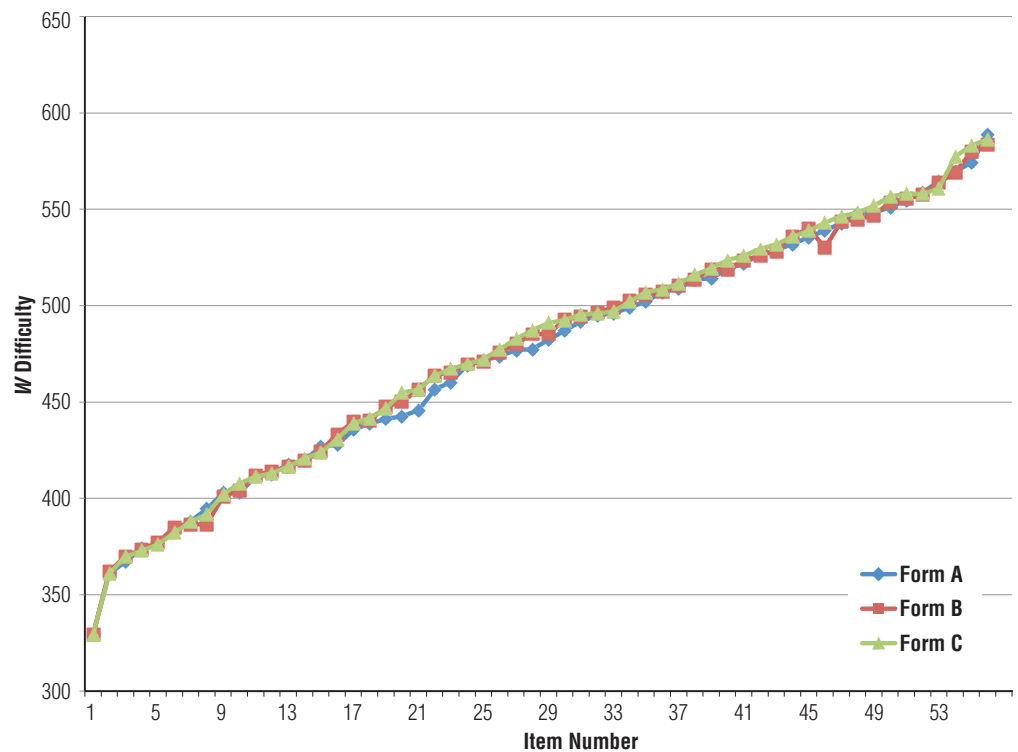


Figure 3.
*Plot of item W difficulties
 across three forms
 of WJ IV ACH Test 3:
 Spelling.*

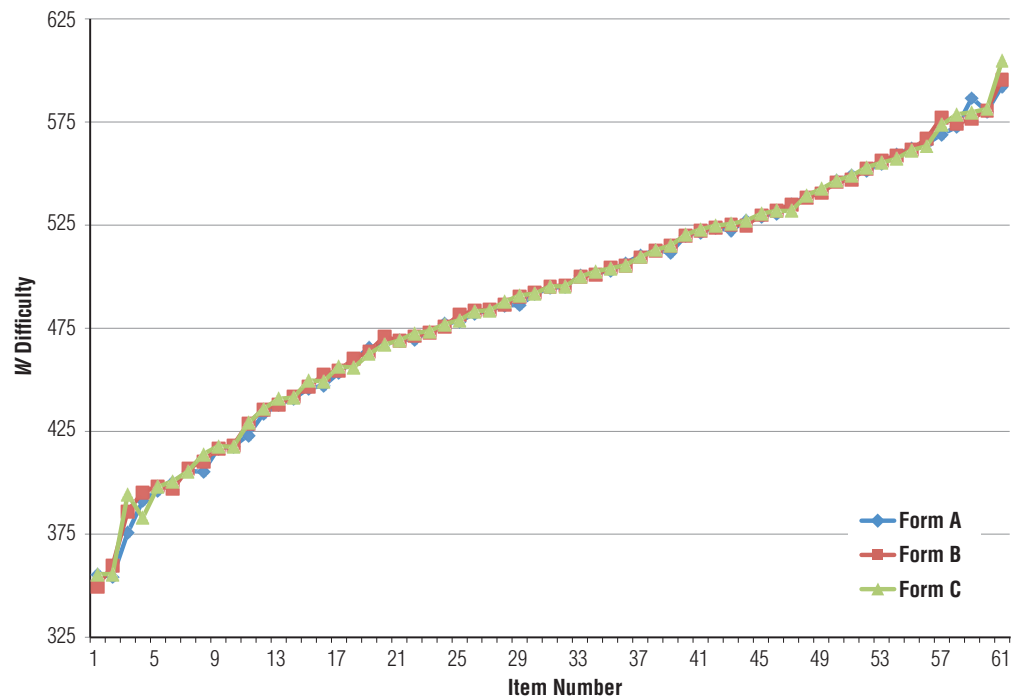


Figure 4.
*Plot of item W difficulties
 across three forms
 of WJ IV ACH Test 4:
 Passage Comprehension.*

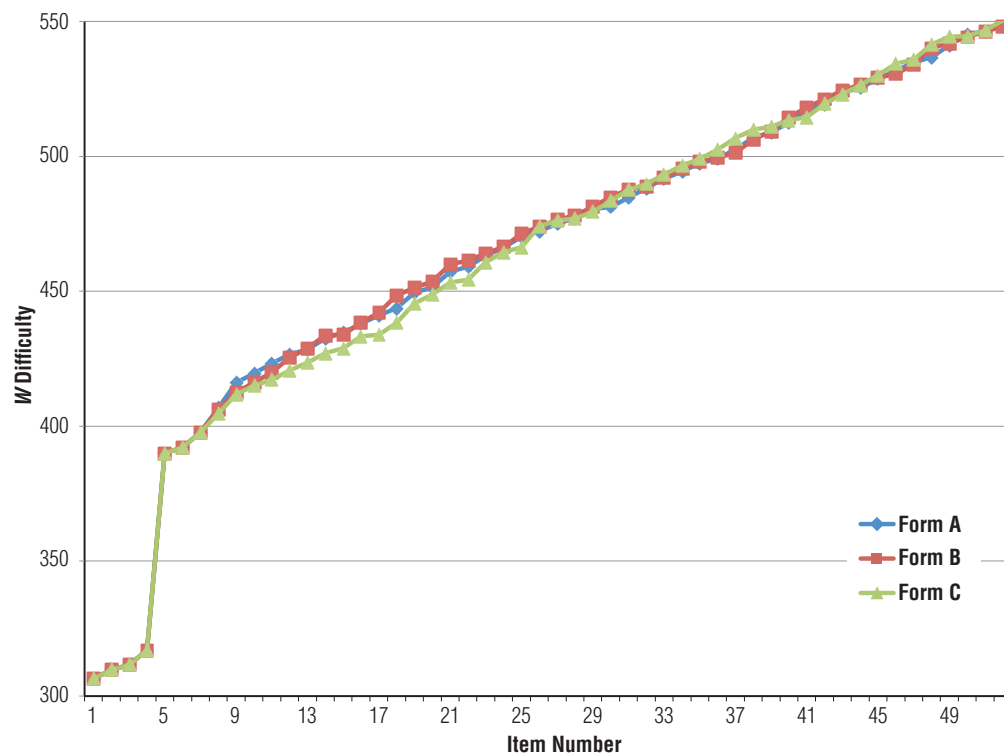


Figure 5.
*Plot of item W difficulties
 across three forms
 of WJ IV ACH Test 5:
 Calculation.*

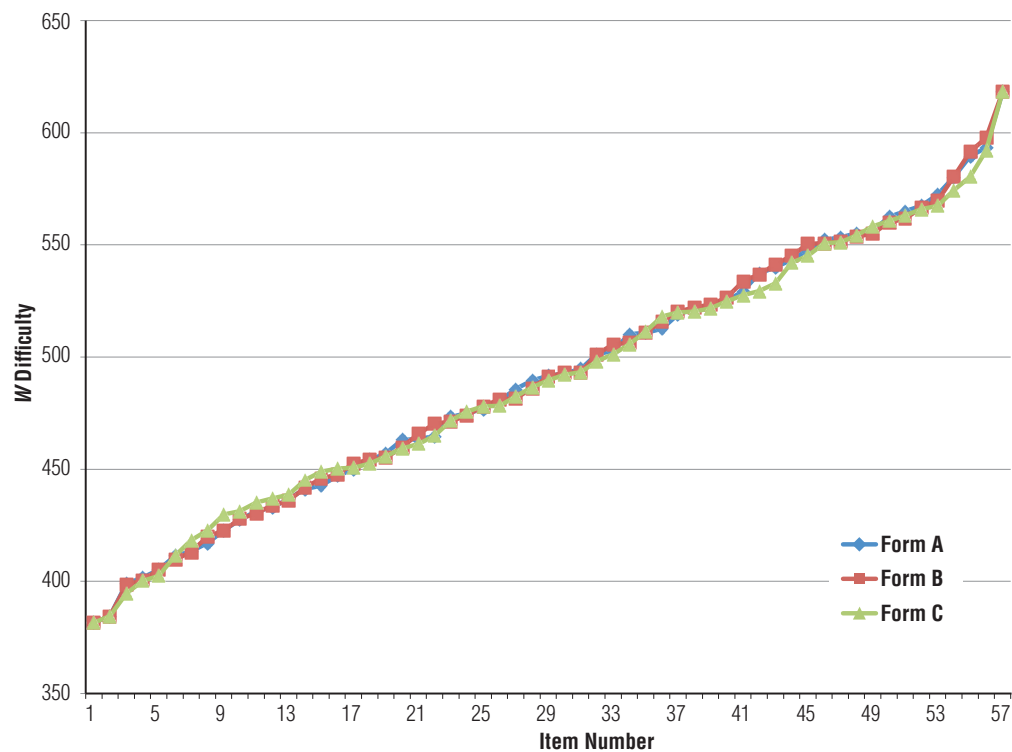


Figure 6.
*Plot of item W difficulties
 across three forms
 of WJ IV ACH Test 6:
 Writing Samples.*

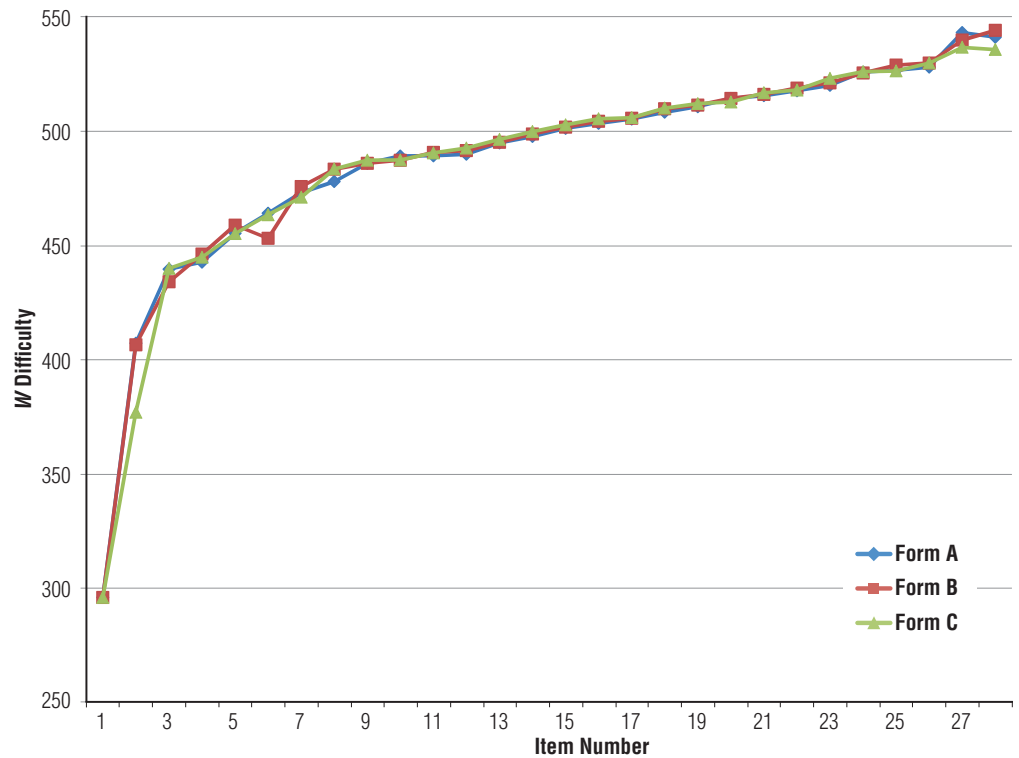


Figure 7.
*Plot of item W difficulties
 across three forms of
 WJ IV ACH Test 7: Word
 Attack.*

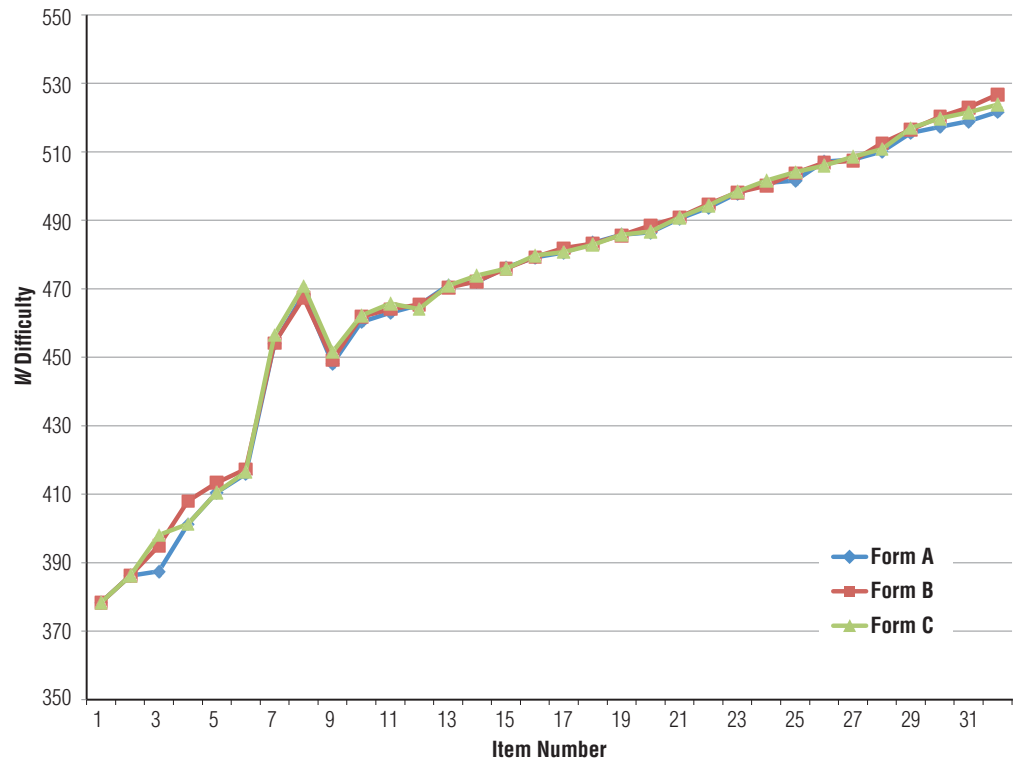
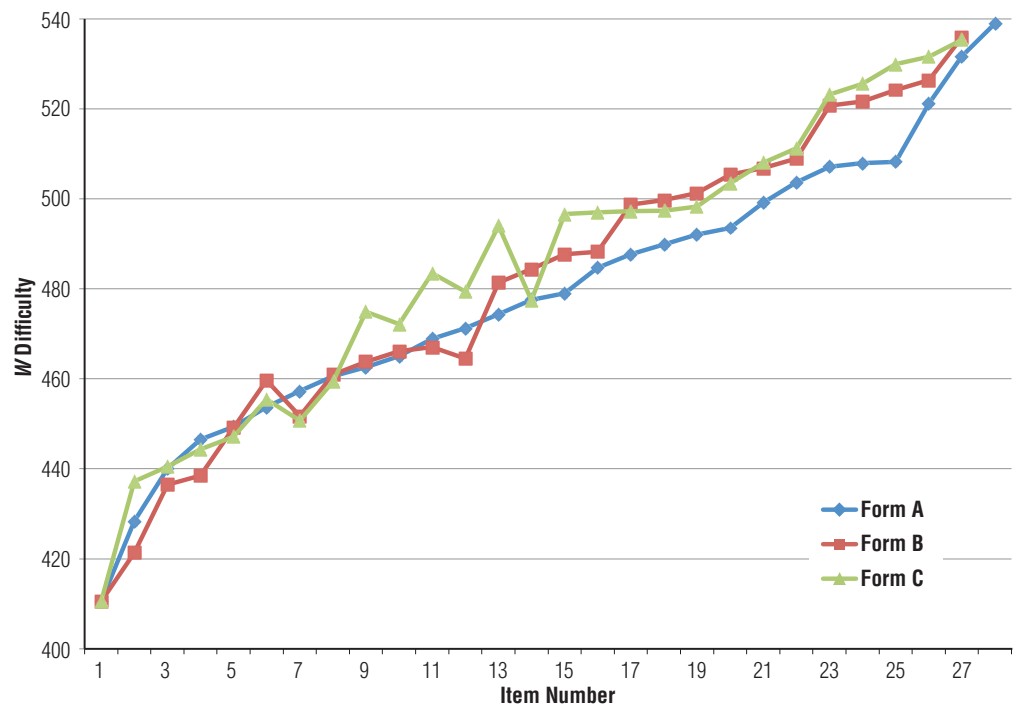


Figure 8.
*Plot of item W difficulties
 across three forms of
 WJ IV ACH Test 8: Oral
 Reading.*



After alternate test forms were constructed through careful item selection, scoring tables were constructed to convert raw test scores into W-ability scores. Because the test forms were constructed from item pools that had been calibrated to a common scale, the scoring tables account for minor differences in item difficulty between forms. The test characteristic curves (TCCs), or raw score-to-W score ogives, provide a visual representation of the relationship between the raw score and the W-ability score for each form of the test. If the same raw score yields a similar W-ability measure across the three forms of the test, and if the same range of W-ability scores is possible on each form, then this provides support for equivalence of test form difficulty. Figures 9 through 16 contain TCC plots showing the relationship of raw score to W ability for each form of the WJ IV ACH Standard Tests 1 through 8. The TCCs are nearly identical for Tests 1 through 7; in places where they differ slightly, the differences can be attributed to small variations in the item difficulties between forms near those locations on the W scale. Examples of these small differences due to item content constraints were discussed earlier. For Test 8: Oral Reading in Figure 16, the raw score-to-W ability ogives are more variable across forms. This variation is due to the item-set-based administration for the Oral Reading test discussed earlier. The equivalence of the TCCs in Figures 9 through 16 provides further evidence that the alternate forms of the WJ IV ACH Standard tests are equivalent in difficulty.

Figure 9.
Relationship between
raw score and W ability
across three forms of
WJ IV ACH Test 1: Letter-
Word Identification.

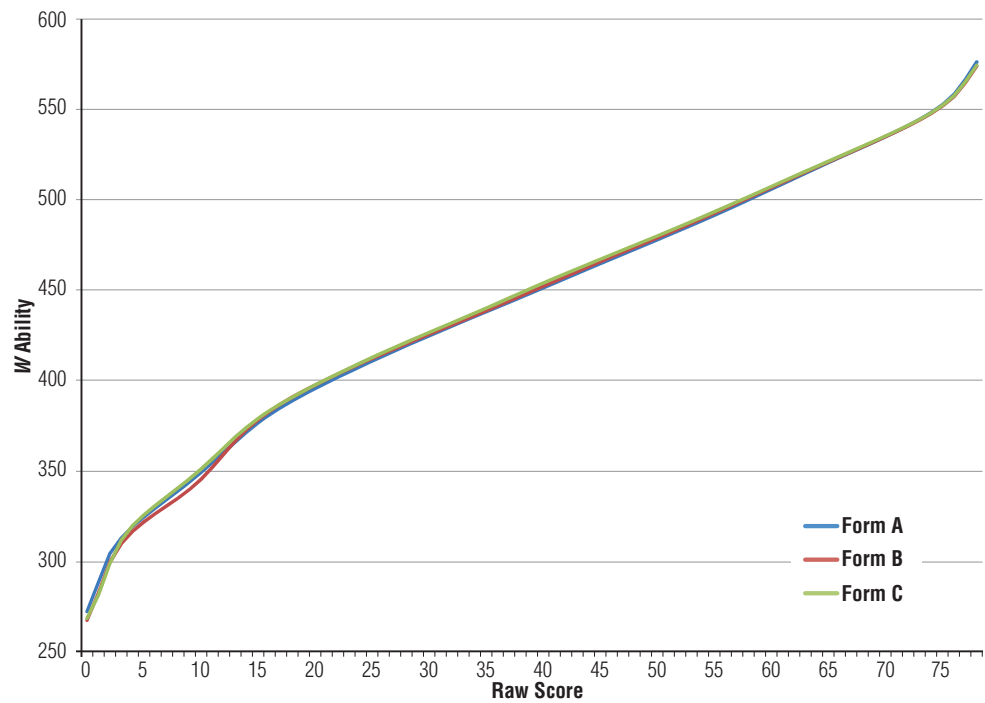


Figure 10.
Relationship between
raw score and W ability
across three forms
of WJ IV ACH Test 2:
Applied Problems.

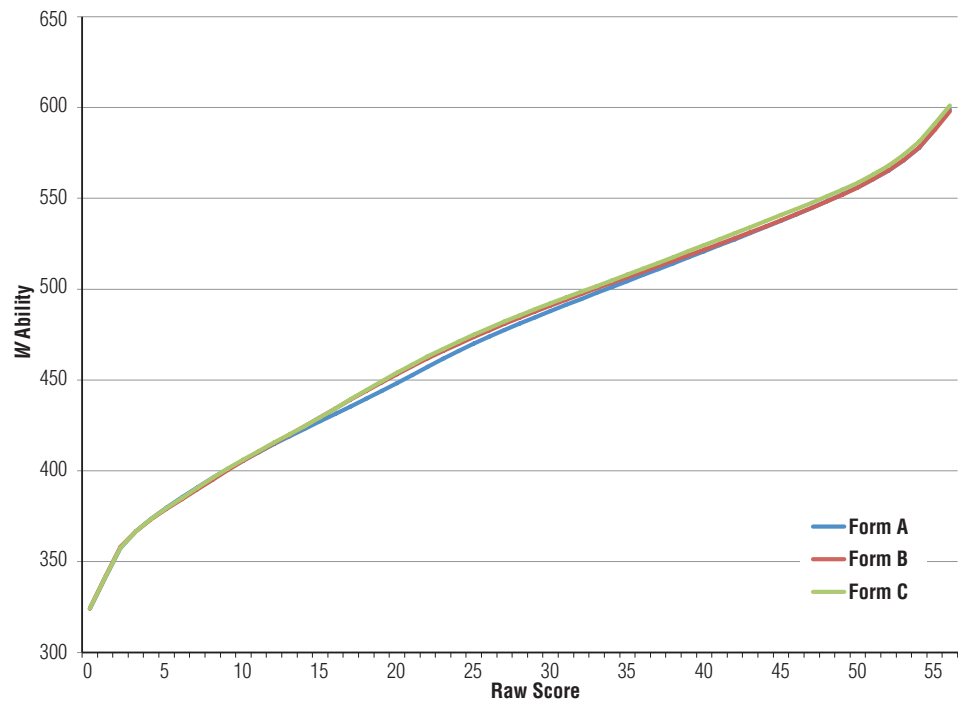


Figure 11.

Relationship between raw score and W ability across three forms of WJ IV ACH Test 3: Spelling.

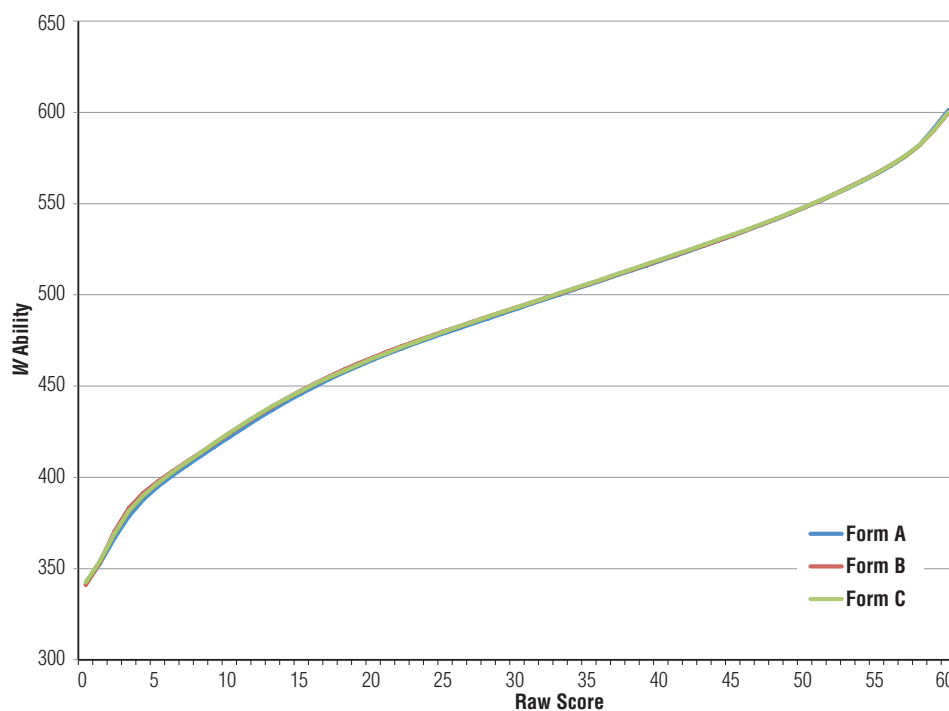


Figure 12.

Relationship between raw score and W ability across three forms of WJ IV ACH Test 4: Passage Comprehension.

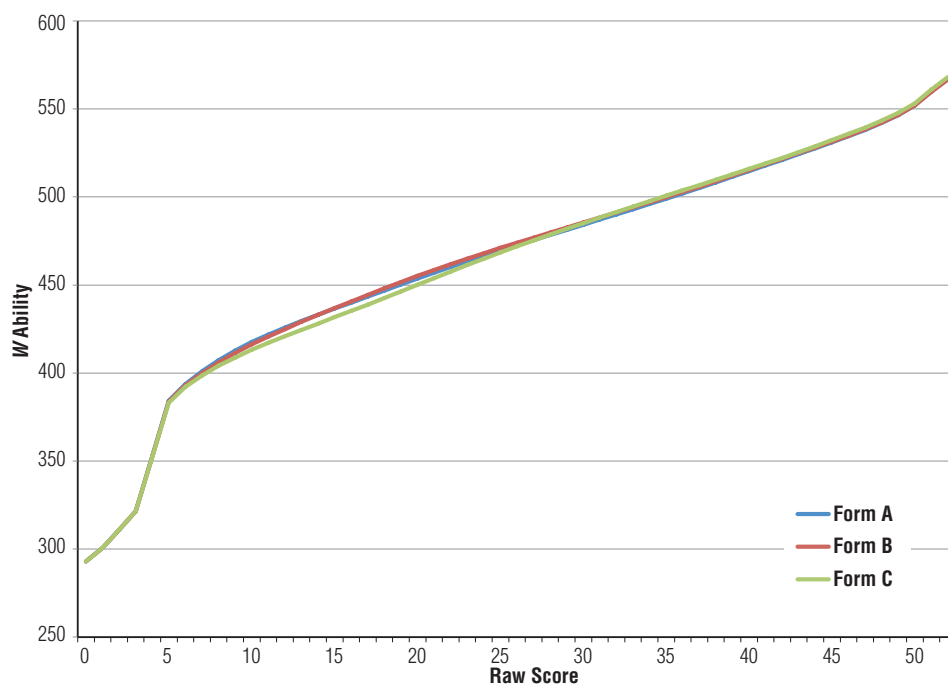


Figure 13.
*Relationship between
raw score and W ability
across three forms
of WJ IV ACH Test 5:
Calculation.*

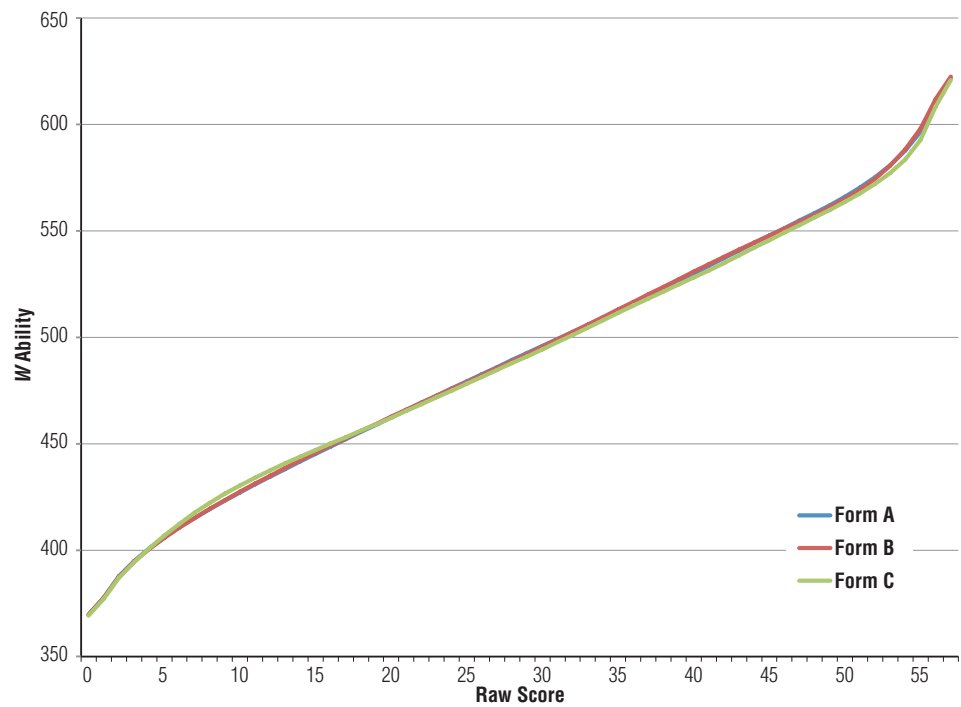


Figure 14.
*Relationship between
raw score and W ability
across three forms
of WJ IV ACH Test 6:
Writing Samples.*

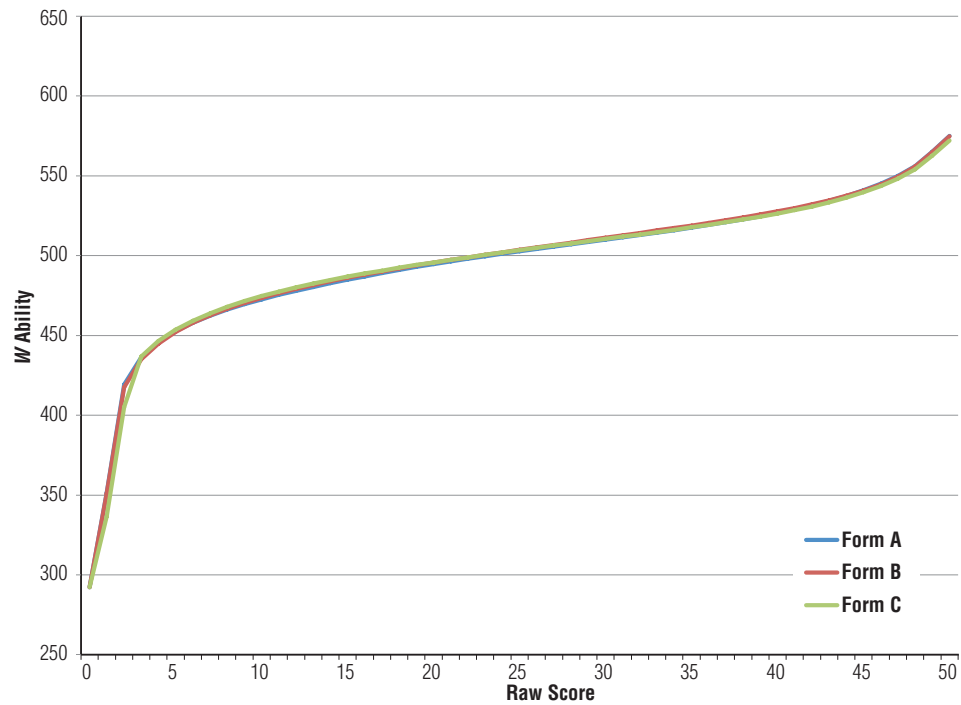


Figure 15.

Relationship between raw score and W ability across three forms of WJ IV ACH Test 7: Word Attack.

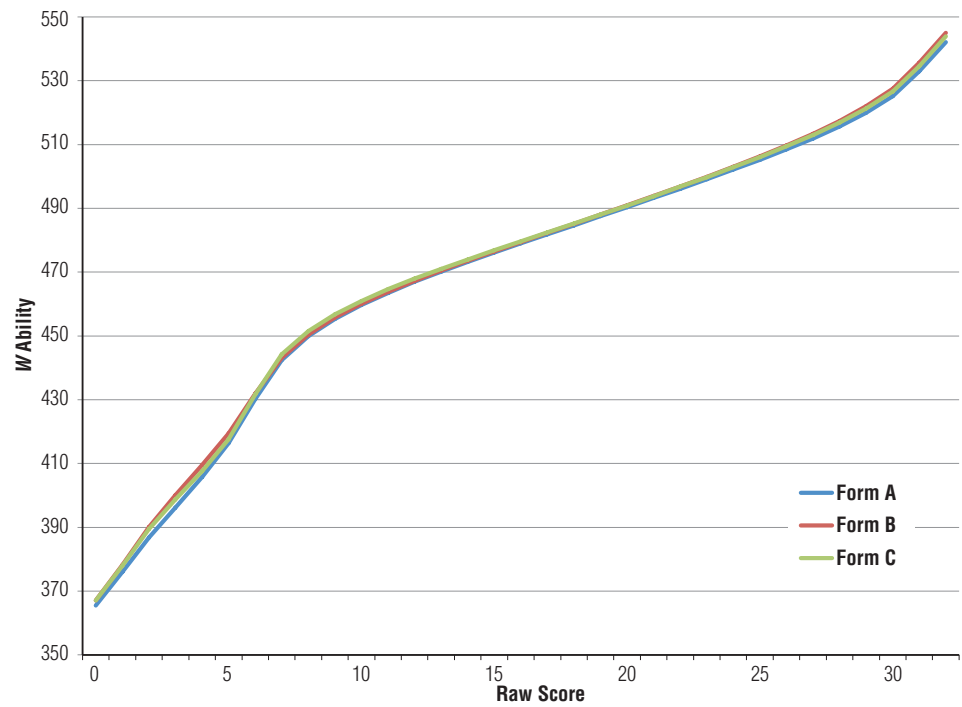
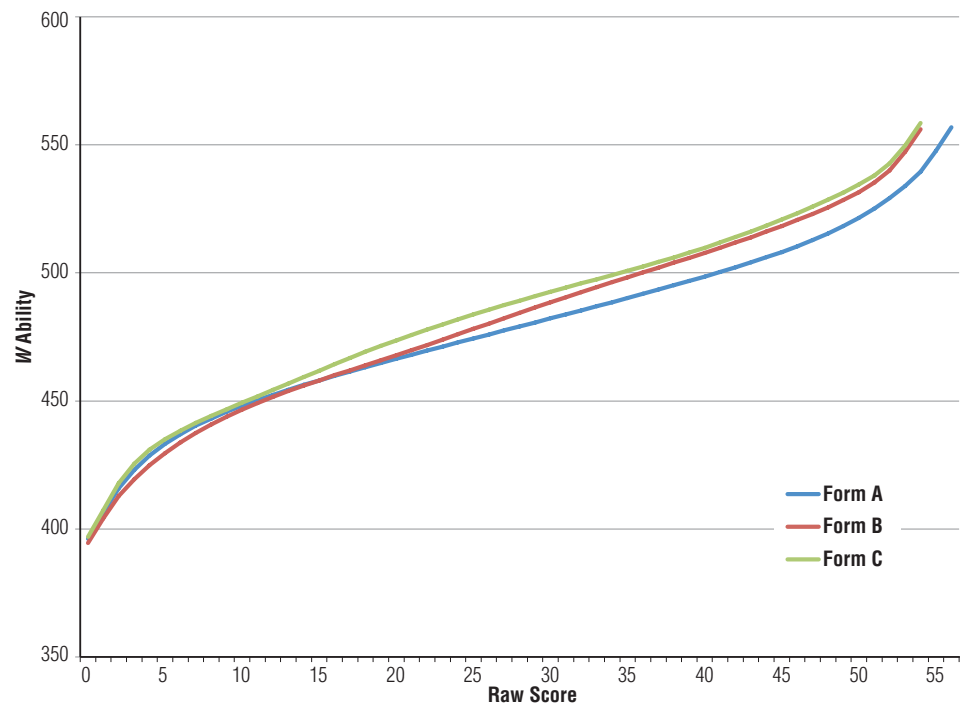


Figure 16.

Relationship between raw score and W ability across three forms of WJ IV ACH Test 8: Oral Reading.



Guideline 3: Equivalence of Measurement Precision

Another means of assessing the equivalence of alternate forms during the test construction stage is to determine whether the forms measure with the same degree of precision across the range of ability. Because there is some degree of error inherent in every measurement, it is best practice to treat test scores as *estimates* of examinees' true ability scores. The standard error of measurement is an index of test score precision, or the accuracy with which the underlying true score can be located on a scale. The *SEM* provides information about the width of the confidence band within which we expect an examinee's observed score to fall, given the examinee's true score. A confidence band of ± 1 *SEM* around the observed score provides an approximate range within which an examinee's true ability score actually lies. Higher *SEM* values result in wider confidence bands, and thus, less measurement precision. Similar *SEM* values across the range of ability for each form of the test provide evidence that examinee scores from alternate forms are equally precise.

Figures 17 through 24 show the relationship between *SEM* and *W* ability for each form of the WJ IV ACH Standard tests. A study of these plots shows that the *SEM*s are comparable across the entire range of ability for most tests. A notable exception is Test 6: Writing Samples. In Figure 22, the *SEM* is notably larger for Forms A and B than for Form C at the lower end of the ability range. This is the effect of the relatively large gap in item difficulty between Items 1 and 2 on Forms A and B (see Figure 6), resulting in less measurement precision at the corresponding location on the *W*-ability scale for those two forms. This between-form difference in measurement precision is a result of the tradeoff between the goal of having unique, but equally difficult, items on each form and the constraints of the limited item pool depth at that particular ability level. This effect is apparent to users as large standard error bands for raw scores of 1 or 2 only on Forms A and B of Test 6: Writing Samples.

Figure 17.
*Relationship between
SEM and W ability
across three forms of
WJ IV ACH Test 1: Letter-
Word Identification.*

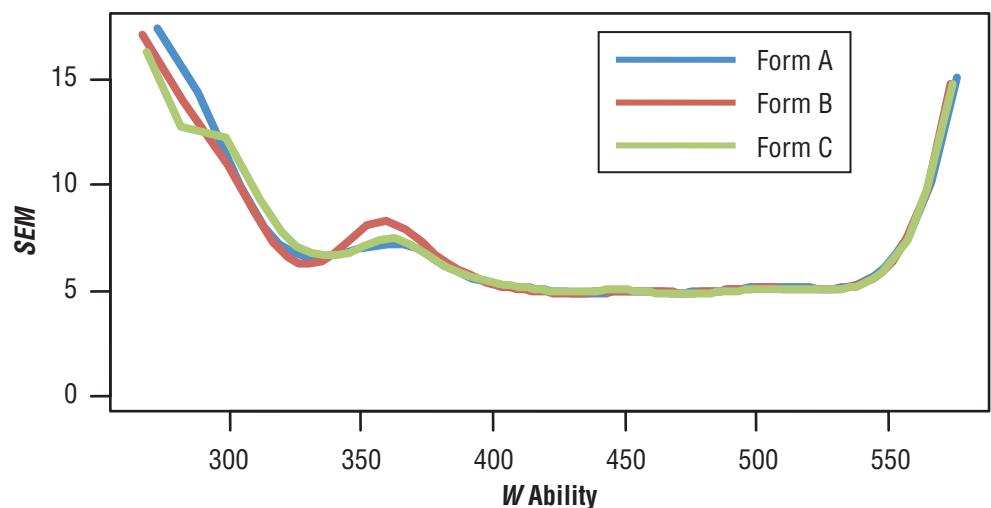


Figure 18.

Relationship between SEM and W ability across three forms of WJ IV ACH Test 2: Applied Problems.

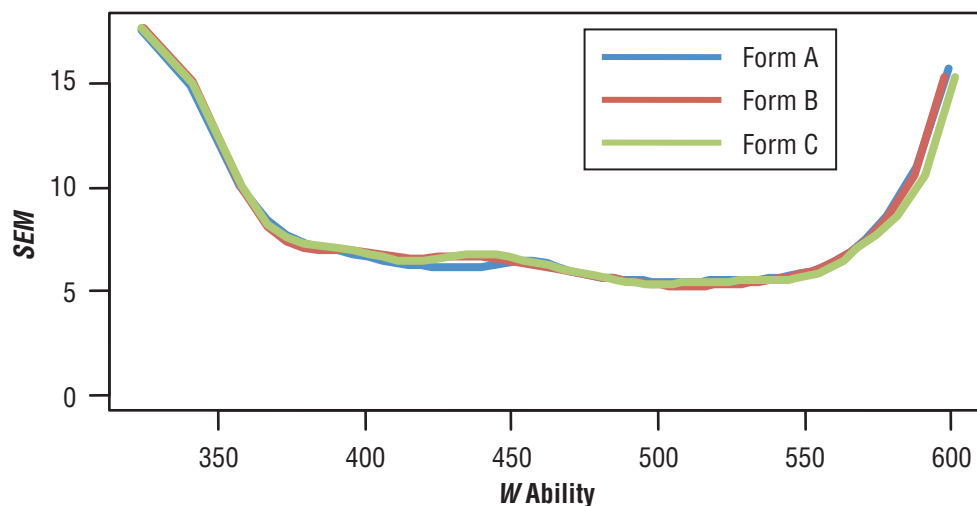


Figure 19.

Relationship between SEM and W ability across three forms of WJ IV ACH Test 3: Spelling.

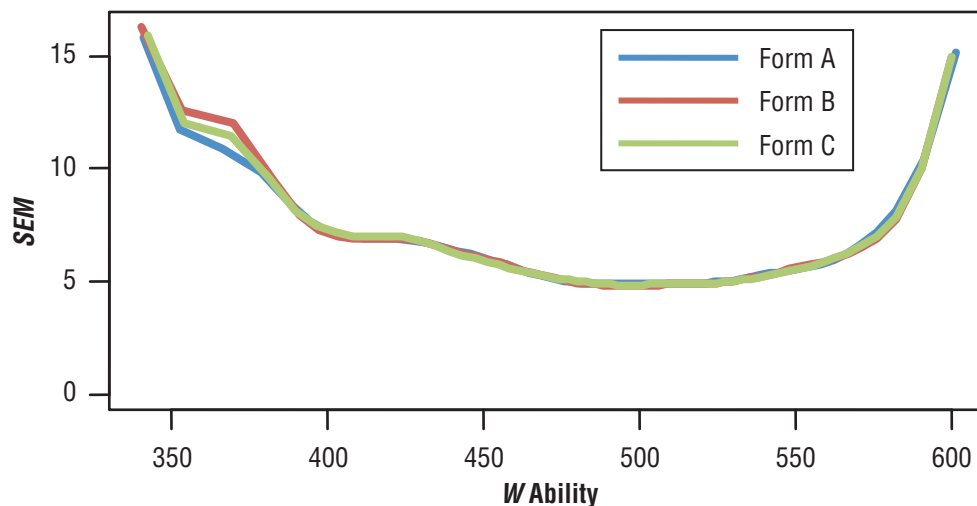


Figure 20.

Relationship between SEM and W ability across three forms of WJ IV ACH Test 4: Passage Comprehension.

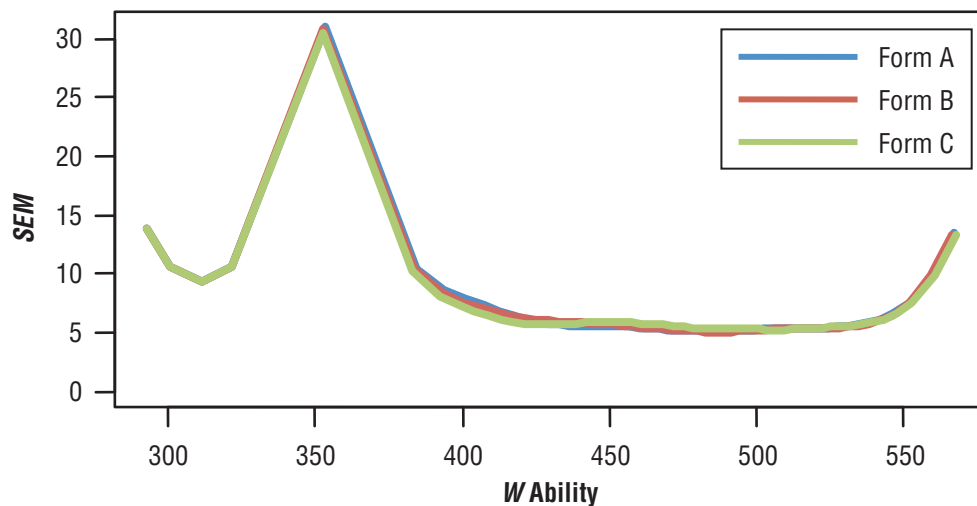


Figure 21.
Relationship between
SEM and W ability
across three forms
of WJ IV ACH Test 5:
Calculation.

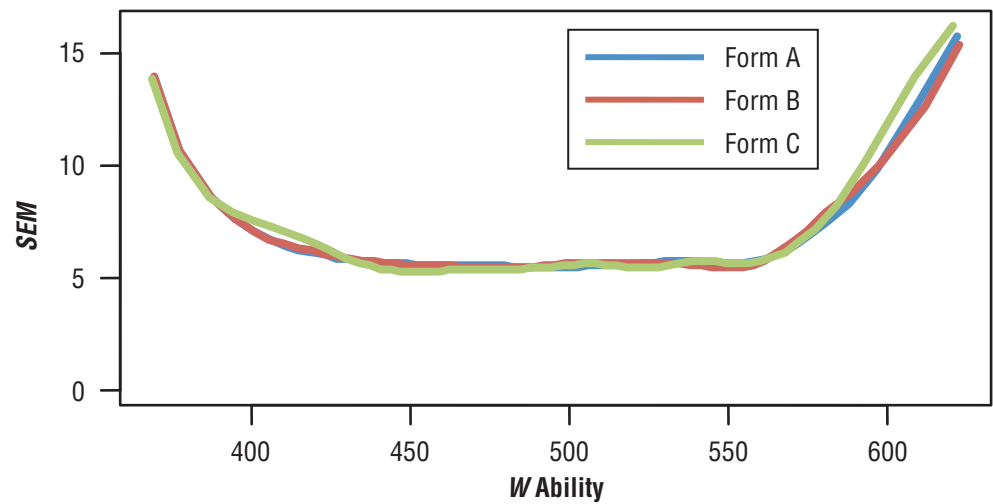


Figure 22.
Relationship between
SEM and W ability
across three forms
of WJ IV ACH Test 6:
Writing Samples.

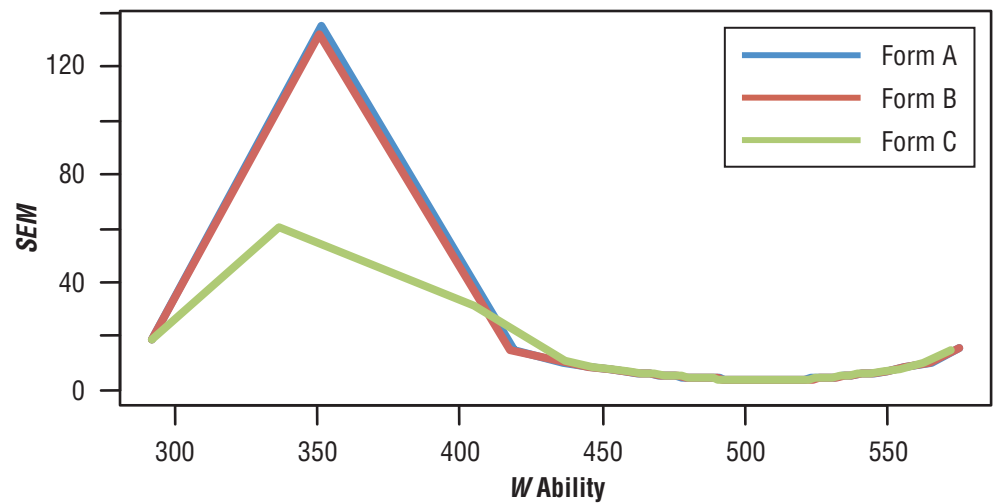


Figure 23.
Relationship between
SEM and W ability
across three forms
of WJ IV ACH Test 7:
Word Attack.

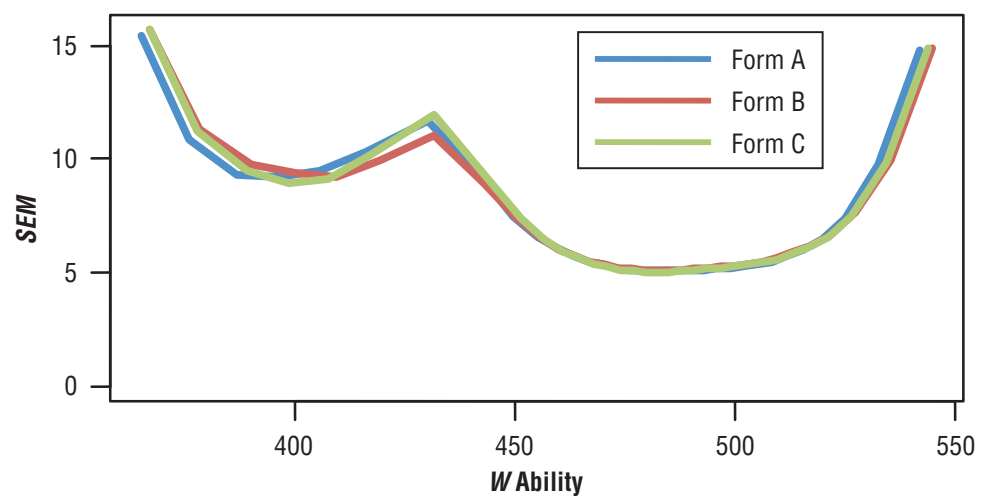
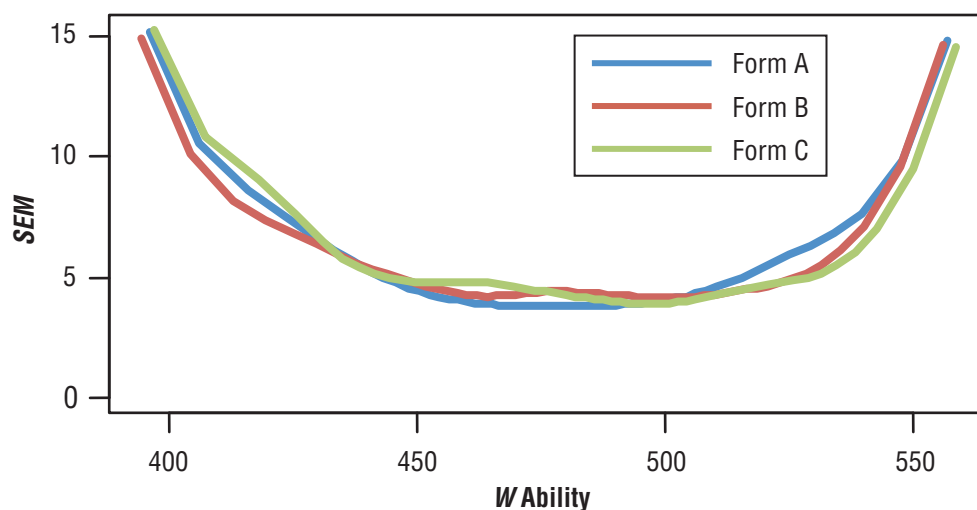


Figure 24.
Relationship between
SEM and W ability
across three forms of
WJ IV ACH Test 8:
Oral Reading.



WJ IV ACH Standard Speeded Tests

All speeded tests in the WJ IV, including the three ACH Standard speeded tests—Test 9: Sentence Reading Fluency, Test 10: Math Facts Fluency, and Test 11: Sentence Writing Fluency—were constructed and calibrated to the W scale using a rate-based metric.² This method of scoring inherently combines information about an examinee’s speed and accuracy of response. In this calibration method, a rate-per-unit-of-time metric was first developed, and all raw scores were converted into the rate-based metric, whereby each examinee had a separate rate score for each minute of testing. These rate scores were then calibrated to the W scale using the Rasch rating scale model (Wright & Masters, 1982), with each minute of testing treated as an “item” in the model. The W difficulty for the first minute of each test—when examinees were typically working most quickly and accurately, resulting in higher rates of correct response—was lower than the second and subsequent minutes of each test—when examinees tended to slow down, become less accurate, or reach the time limit, resulting in a lower rate of correct response. In operational use, the rate-based method requires as input only a total count of correct responses and a total testing time.

The rate-based scoring method makes two assumptions about the difficulty of the items on the test that are relevant to the examination of alternate-forms equivalence. First, the average difficulty of the item types on each test is more relevant to the overall form difficulty than is the difficulty of any individual item. Second, it follows that for alternate forms of speeded tests, it is more important for the average difficulty of the items on each form to be equivalent, rather than for there to be a one-to-one correspondence of item difficulty by serial item position (as on the nonspeeded alternate forms of the WJ IV ACH Standard tests). Although care was taken to ensure that item types shifted from easy to more difficult at the same serial positions on all three forms of each test, item difficulty does vary by item within each row or page of the test stimulus materials for these tests. For example, all three forms of Test 10: Math Facts Fluency contain 16 rows of math facts problems that an examinee must complete as quickly as possible. Although the exact content of the items varies between forms, the forms were constructed so that the approximate difficulty of the items in each corresponding row of items across the three forms is approximately equal.

² For detailed information about the derivation of the rate-based metric for calibration of the WJ IV speeded tests, see the “Calibration of Timed Tests” section in Chapter 2 of the *WJ IV Technical Manual* (McGrew, LaForte, & Schrank, 2014).

In summary, for both the speeded and nonspeeded tests in the WJ IV ACH Standard Battery, great care was taken during the test design, item calibration and equating, and form construction stages to ensure that the three forms of each test would contain equivalent content coverage, would be equally difficult for all examinees, and would provide the same range of measurement precision across the entire ability range. These goals were generally met for all tests; in cases where item difficulty calibrations or measurement error varies slightly between forms, these variations are the result of limitations due to the item type or item pool depth at specific ability levels.

Part B: Alternate-Forms Equivalence Studies

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) state, “In theory, equating *should* provide accurate score conversions for any set of persons drawn from the examinee population for which the test is designed” (p. 51, emphasis added). In practice, however, there are many factors that could affect the operational equivalence of test forms for which even the most careful equating procedures were employed. These factors include variations in examinee motivation across administrations, item difficulty “drift” over time, and curricular shifts over time, to name a few. For this reason, it is important that the body of evidence to support alternate-forms equivalence for any test includes results from empirical studies conducted with several different samples and in different settings.

Because most individually administered achievement tests are not available in multiple forms, few current studies examine operational form equivalence of these types of tests. Several studies have examined the equivalence of alternate forms for cognitive or neuropsychological assessments (e.g., Hinton-Bayre & Geffen, 2005; McCallum & Bracken, 1981; McCarty, Ziesat, Logue, Power, & Rosenstiel, 1980; Ross, Furr, Carter, & Weinberg, 2006; Ryan, Geisser, Randall, & Georgemiller, 1986; and Schmidt, Mattis, Adams, & Nestor, 2005). Many studies report only Pearson product-moment correlations between scores from two or more forms, termed “alternate-forms reliability coefficients” or “coefficients of equivalence.” While these reliability coefficients are useful in predicting the consistency of an examinee’s rank order in a sample where the examinees have been administered two or more forms of a test, they fail to provide information about relative form difficulty. If study results are to be useful in assuring test users that scores from alternate forms of a test are comparable, the studies must evaluate not only the reliability of the alternate forms but also the equivalence of the form difficulties.

At the conclusion of the WJ IV norming study, two studies were conducted to assess the operational equivalence of the three publication forms of each of the WJ IV ACH Standard Battery tests. The first study included the nonspeeded tests comprising the following WJ IV clusters: Reading (Test 1: Letter-Word Identification and Test 4: Passage Comprehension), Mathematics (Test 2: Applied Problems and Test 5: Calculation), and Written Language (Test 3: Spelling and Test 6: Writing Samples). The second study included the speeded tests that compose the WJ IV Academic Fluency cluster (Test 9: Sentence Reading Fluency, Test 10: Math Facts Fluency, and Test 11: Sentence Writing Fluency).

Nonspeeded Tests Alternate-Forms Equivalence Study

Data for the WJ IV ACH Standard nonspeeded tests alternate-forms study (AFS) were gathered from January through March 2014. Examiners who had been hired and trained by Riverside recruited convenience samples of study participants in the following three groups: grades 3 through 4, grades 9 through 12, and college+. Examinees could not have participated in a prior Woodcock-Johnson IV research study. Individuals with a diagnosis of cognitive or developmental delay and individuals with IEPs or 504 plans were not included in the studies. After each participant (or their parent or guardian) completed a consent form, the examiner randomly assigned the individual to an administration condition within either the Reading (Study 1A), Mathematics (Study 1B), or Written Language (Study 1C) alternate-forms study. Administration conditions counterbalanced the order of test and form administration within each study. Table 2 contains the total number and distribution of examinees in each administration condition for the WJ IV ACH Standard nonspeeded alternate-forms study.

Table 2.
*Distribution of
Examinees in Each
Administration Condition
for the WJ IV ACH
Standard Nonspeeded
Tests Alternate-Forms
Study*

	Study Group	<i>n</i>	First Test	Administration Order	Second Test	Administration Order
Study 1A: WJ IV Reading Cluster Tests (<i>N</i> = 121)	Grades 3–4	20	Test 1: Letter-Word Identification	C-B-A	Test 4: Passage Comprehension	C-B-A
	Grades 9–12	20	Test 1: Letter-Word Identification	A-C-B	Test 4: Passage Comprehension	A-C-B
	College+	20	Test 1: Letter-Word Identification	B-C-A	Test 4: Passage Comprehension	B-C-A
	Grades 3–4	21	Test 4: Passage Comprehension	A-B-C	Test 1: Letter-Word Identification	A-B-C
	Grades 9–12	20	Test 4: Passage Comprehension	B-A-C	Test 1: Letter-Word Identification	B-A-C
	College+	20	Test 4: Passage Comprehension	C-A-B	Test 1: Letter-Word Identification	C-A-B
Study 1B: WJ IV Mathematics Cluster Tests (<i>N</i> = 119)	Grades 3–4	20	Test 2: Applied Problems	A-C-B	Test 5: Calculation	A-C-B
	Grades 9–12	20	Test 2: Applied Problems	B-C-A	Test 5: Calculation	B-C-A
	College+	19	Test 2: Applied Problems	C-B-A	Test 5: Calculation	C-B-A
	Grades 3–4	20	Test 5: Calculation	B-A-C	Test 2: Applied Problems	B-A-C
	Grades 9–12	20	Test 5: Calculation	A-B-C	Test 2: Applied Problems	A-B-C
	College+	20	Test 5: Calculation	C-A-B	Test 2: Applied Problems	C-A-B

Table 2. (cont.)
Distribution of
Examinees in Each
Administration Condition
for the WJ IV ACH
Standard Nonspeeded
Tests Alternate-Forms
Study

	Study Group	<i>n</i>	First Test	Administration Order	Second Test	Administration Order
Study 1C: WJ IV Written Language Cluster Tests (<i>N</i> = 120)	Grades 3–4	20	Test 3: Spelling	C-B-A	Test 6: Writing Samples	A-B
	Grades 9–12	20	Test 3: Spelling	B-C-A	Test 6: Writing Samples	A-C
	College+	20	Test 3: Spelling	A-C-B	Test 6: Writing Samples	B-C
	Grades 3–4	20	Test 6: Writing Samples	C-B	Test 3: Spelling	C-A-B
	Grades 9–12	20	Test 6: Writing Samples	B-A	Test 3: Spelling	B-A-C
	College+	20	Test 6: Writing Samples	C-A	Test 3: Spelling	A-B-C

Note. Examinees in Study 1C were each administered only two forms of Writing Samples to reduce the likelihood of examinee attrition due to fatigue.

Table 3 contains demographic information about the samples for the WJ IV ACH Standard nonspeeded tests alternate-forms study. Because study participant groups were samples of convenience, demographic characteristics are not necessarily representative of the U.S. population. Within each study group, participants predominantly are White and Not Hispanic. For the grades 3 through 4 and grades 9 through 12 study groups, the level of parent education generally is higher than in the average U.S. population. For the college+ sample, the level of educational attainment also is significantly higher than in the average U.S. population. For example, 74% of alternate-forms study participants in this group have at least a Bachelor's degree, while current U.S. Census figures report that 28.5% of individuals ages 25 and older in the United States have at least a Bachelor's degree (U.S. Census Bureau, 2014).

Table 3.
Sample Demographic
Characteristics for the
WJ IV ACH Standard
Nonspeeded Tests
Alternate-Forms Study

Demographic Characteristics	Study Group		
	Grades 3–4	Grades 9–12	College+
	<i>N</i> = 121	<i>N</i> = 120	<i>N</i> = 119
Age (years)			
Range	8–10	14–18	18–66
Mean	9.4	16.2	39.1
<i>SD</i>	0.5	1.2	8.4
Sex			
Male	57.0	40.8	26.9
Female	43.0	59.2	73.1
Race			
White	86.8	97.5	97.5
Black	7.4	0.8	1.7
American Indian/Alaska Native	—	—	—
Asian/Native Hawaiian/Pacific Islander	5.0	1.7	0.4
Other/Mixed	0.8	—	—

Table 3. (cont.)
*Sample Demographic
 Characteristics for the
 WJ IV ACH Standard
 Nonspeeded Tests
 Alternate-Forms Study*

Demographic Characteristics	Study Group		
	Grades 3–4	Grades 9–12	College+
	<i>N</i> = 121	<i>N</i> = 120	<i>N</i> = 119
Ethnicity			
Not Hispanic	91.7	93.3	94.1
Hispanic	8.3	6.7	5.9
Mother's Education			
No Information Provided	—	8.3	—
< High School Graduate	—	—	—
High School Graduate	10.7	19.2	—
> High School	89.3	72.5	—
Father's Education			
No Information Provided	1.7	8.3	—
< High School Graduate	1.7	5.8	—
High School Graduate	19.8	23.3	—
> High School	76.9	62.5	—
Examinee's Grade in School			
Grade 3	38.8	—	—
Grade 4	61.2	—	—
Grade 9	—	37.5	—
Grade 10	—	18.3	—
Grade 11	—	22.5	—
Grade 12	—	21.7	—
Examinee's Highest Year of Education Completed			
First Year of College	—	—	13.4
Second Year of College	—	—	10.1
Third Year of College	—	—	2.5
Bachelor's Degree	—	—	38.7
Master's Degree or Higher	—	—	35.3

Table 4 contains sample sizes, mean W-ability scores, and standard deviations for Forms A, B, and C of all six WJ IV ACH Standard nonspeeded tests, by study group. W scores, rather than standard scores, are reported for this analysis because W scores are an interval unit of measurement and do not depend on the underlying distribution of examinees from the WJ IV norming sample for interpretation. When assessing the equivalence of form difficulties, the primary question of interest is: “Are examinees who are administered a particular form of the test at a disadvantage when compared to examinees who are administered another form of the test?” Similar mean observed W-ability scores on all three forms provide evidence that the form difficulties are equivalent. If these mean score similarities are replicated across multiple samples, the evidence for form equivalence is stronger.

Table 4.
Sample Sizes, Mean
W-Ability Scores, and
Standard Deviations
for Each Form of Six
WJ IV ACH Standard
Nonspeeded Tests

Study Group	Form	Study 1A: Reading					
		Test 1: Letter-Word Identification			Test 4: Passage Comprehension		
		<i>n</i>	Mean <i>W</i>	<i>SD</i>	<i>n</i>	Mean <i>W</i>	<i>SD</i>
Grades 3–4	A	41	499.8	16.8	41	491.8	12.8
	B	41	499.3	18.5	41	494.0	11.6
	C	41	501.1	15.0	41	491.3	12.2
Grades 9–12	A	40	528.4	12.8	40	521.9	12.2
	B	40	530.6	13.1	40	519.4	10.9
	C	40	531.3	12.4	40	517.6	11.4
College+	A	40	548.4	14.5	40	534.8	9.8
	B	40	547.6	12.1	40	538.0	12.3
	C	40	543.0	11.4	40	527.3	12.2

Study Group	Form	Study 1B: Mathematics					
		Test 2: Applied Problems			Test 5: Calculation		
		<i>n</i>	Mean <i>W</i>	<i>SD</i>	<i>n</i>	Mean <i>W</i>	<i>SD</i>
Grades 3–4	A	41	506.2	13.2	41	487.8	12.4
	B	41	500.4	13.6	41	492.6	13.8
	C	41	501.7	12.7	41	492.3	12.4
Grades 9–12	A	40	534.7	13.5	40	530.6	15.1
	B	40	528.3	14.8	40	532.7	15.8
	C	40	529.2	14.8	40	530.4	15.3
College+	A	39	537.2	16.1	39	530.3	17.6
	B	39	533.4	17.7	39	530.2	15.4
	C	39	536.1	17.1	39	529.5	15.8

Study Group	Form	Study 1C: Written Language					
		Test 3: Spelling			Test 6: Writing Samples ^a		
		<i>n</i>	Mean <i>W</i>	<i>SD</i>	<i>n</i>	Mean <i>W</i>	<i>SD</i>
Grades 3–4	A	41	499.4	16.5	20	499.0	8.0
	B	41	497.1	16.1	20 ^b	495.0	7.2
					20 ^c	489.9	8.4
	C	41	503.1	15.4	20	492.3	11.6
Grades 9–12	A	40	531.6	12.2	20 ^b	518.6	6.8
					20 ^d	514.8	11.0
	B	40	531.6	12.5	20	514.3	6.3
	C	40	538.9	12.5	20	518.3	10.0
College+	A	40	545.1	13.2	20	527.9	10.6
	B	40	544.3	13.5	20	522.5	7.5
	C	40	552.4	13.3	20 ^c	521.7	8.4
					20 ^d	523.1	13.8

^a The Writing Samples scores in the table are broken out by matched sample for Form B in the grades 3–4 sample, Form A in the grades 9–12 sample, and Form C in the college+ sample. For all 40 examinees who were administered Form B in the grades 3–4 sample, the mean score was 492.5 and the *SD* was 8.1. For all 40 examinees who were administered Form A in the grades 9–12 sample, the mean score was 516.7 and the *SD* was 9.3. For all 40 examinees who were administered Form C in the college+ sample, the mean score was 522.4 and the *SD* was 11.3.

^b Examinees were administered Forms A and B only.

^c Examinees were administered Forms B and C only.

^d Examinees were administered Forms A and C only.

Table 5 contains the results of *t* tests from all possible pairwise comparisons of the mean *W* scores that were presented for each test and study group in Table 4. The results presented in Tables 4 and 5 taken together provide information about whether *statistically significant* differences exist between mean *W*-ability scores on different forms of the tests. For the grades 3 through 4 sample, the mean examinee *W* score on Form A of Applied Problems (506.2) was significantly higher than the mean examinee *W* scores on Forms B (500.4) and C (501.7), the mean examinee *W* score on Form A of Calculation (487.8) was significantly lower than the mean examinee *W* scores on Forms B (492.6) and C (492.3), and the mean examinee *W* score on Form C of Spelling (503.1) was significantly higher than the mean examinee *W* scores on Forms A (499.4) and B (497.1). For the grades 9 through 12 sample, the mean examinee *W* score on Form C of Passage Comprehension (517.6) was significantly lower than the mean examinee *W* score on Form A (521.9) but was comparable to the mean examinee *W* score on Form B (519.4), the mean examinee *W* score on Form A of Applied Problems (534.7) was significantly higher than the mean examinee *W* score on Forms B (528.3) and C (529.2), and the mean examinee *W* score on Form C of Spelling (538.9) was significantly higher than the mean examinee *W* scores on Forms A (531.6) and B (531.6). For the college+ sample, the mean examinee *W* score on Form C of Letter-Word Identification (543.0) was significantly lower than the mean examinee *W* score on Forms A (548.4) and B (547.6), the mean examinee *W* score on Form C of Passage Comprehension (527.3) was significantly lower than the mean examinee *W* scores on Forms A (534.8) and B (538.0), and the mean examinee *W* score on Form C of Spelling (552.4) was significantly higher than the mean examinee *W* scores on Forms A (545.1) and B (544.3).

Table 5.
T-Test Results From
Pairwise Comparisons
of Alternate Forms of
the WJ IV ACH Standard
Nonspeeded Tests

Study Group	Form Comparison	Study 1A: Reading					
		Test 1: Letter-Word Identification			Test 4: Passage Comprehension		
		<i>t</i>	<i>df</i>	<i>p</i>	<i>t</i>	<i>df</i>	<i>p</i>
Grades 3–4	A & B	0.335	40	1.000	-1.627	40	0.335
	A & C	-1.298	40	0.605	0.307	40	1.000
	B & C	-1.565	40	0.377	2.447	40	0.057
Grades 9–12	A & B	-1.566	39	0.376	1.646	39	0.324
	A & C	-2.242	39	0.092	3.434	39	0.004
	B & C	-0.555	39	1.000	1.153	39	0.768
College+	A & B	0.561	39	1.000	-2.442	39	0.058
	A & C	3.537	39	0.003	6.485	39	< 0.001
	B & C	3.373	39	0.005	8.174	39	< 0.001

Study Group	Form Comparison	Study 1B: Mathematics					
		Test 2: Applied Problems			Test 5: Calculation		
		<i>t</i>	<i>df</i>	<i>p</i>	<i>t</i>	<i>df</i>	<i>p</i>
Grades 3–4	A & B	4.314	39	< 0.001	-4.455	39	< 0.001
	A & C	3.709	39	0.002	-5	39	< 0.001
	B & C	-1.119	39	0.810	0.291	39	1.000
Grades 9–12	A & B	5.170	39	< 0.001	-1.777	39	0.250
	A & C	4.317	39	< 0.001	0.132	39	1.000
	B & C	-0.765	39	1.000	1.667	39	0.310
College+	A & B	2.783	38	0.025	0.067	38	1.000
	A & C	0.746	38	1.000	0.626	38	1.000
	B & C	-2.38	38	0.067	0.564	38	1.000

Table 5. (cont.)
T-Test Results From
Pairwise Comparisons
of Alternate Forms of
the WJ IV ACH Standard
Nonspeeded Tests

Study Group	Form Comparison	Study 1C: Written Language					
		Test 3: Spelling			Test 6: Writing Samples		
		<i>t</i>	<i>df</i>	<i>p</i>	<i>t</i>	<i>df</i>	<i>p</i>
Grades 3–4	A & B	2.034	39	0.146	1.756	19	0.190
	A & C	–4.029	39	0.001	—	—	—
	B & C	–5.783	39	< 0.001	–1.535	19	0.282
Grades 9–12	A & B	0.010	39	1.000	2.518	19	0.042
	A & C	–5.19	39	< 0.001	–2.403	19	0.054
	B & C	–6.179	39	< 0.001	—	—	—
College+	A & B	0.722	39	1.000	—	—	—
	A & C	–5.285	39	< 0.001	2.600	19	0.036
	B & C	–6.784	39	< 0.001	0.621	19	1.000

Note. Bold values indicate that the pair of mean scores reported in Table 4 are significantly different ($p < .05$). Bonferroni correction for multiple comparisons was applied to p values.

For the Writing Samples test, the time constraints of test administration allowed for only two forms of the test to be administered to each examinee. As a result, the sample sizes for each of the paired comparisons are $n = 20$. For the examinees in the grades 3 through 4 sample, there were no significant differences between the mean scores for examinees who were administered either Writing Samples Forms A and B or Writing Samples Forms B and C. For the grades 9 through 12 sample, examinees who were administered Writing Samples Forms A and B scored significantly higher on Form A (mean = 518.6) than on Form B (mean = 514.3). Performance on Writing Samples Forms A and C was similar for examinees who were administered those two forms. For the college+ sample, the examinees who were administered Writing Samples Forms A and C scored significantly higher on Form A (mean = 527.9) than on Form C (mean = 523.1), but there was no significant difference between scores on Writing Samples Forms B and C for examinees who were administered those two forms.

The t tests reported in Table 5 provide information about the *statistical significance* of mean W-score differences between forms within each study group that are reported in Table 4. To determine whether these mean score differences have *practical meaning*, it is useful to examine the score differences in the context of the measurement error associated with each examinee's observed score. Using the SEM associated with each observed score, the high and low W-ability scores marking the bounds of the 68% confidence bands around each examinee's score were computed. These bands were then compared to see whether they overlapped. When an examinee's W-score confidence bands from alternate forms do not overlap, a meaningful difference might exist between the examinee's W scores on the two forms.

Table 6 displays the percentage of examinees in each study group that had overlapping 68% confidence bands for each pair of forms administered (i.e., ± 1 SEM around the examinee's observed scores). High percentages (i.e., 80% and higher) of overlapping examinee score confidence bands for a pair of test forms suggests that any apparent differences in the mean observed examinee scores between forms might be due to measurement error in individual examinee scores, rather than a difference in the mean true scores of examinees and thus, the difficulty of the forms. Evaluating the information about the mean score differences presented in Tables 4 and 5, together with the percentages in Table 6, may assist users in determining whether an observed mean score difference is practically meaningful, or whether it is more likely due to expected

variations from random measurement error. For example, in Table 6 the percentage of examinees in the college+ study group with overlapping W-score confidence bands for Test 4: Passage Comprehension is relatively low for form pairs A and C (62.5%) and form pairs B and C (60%). These two form pairs also exhibited statistically significant mean score differences for this sample of examinees (see Tables 4 and 5). Taking this evidence together, users might conclude that the examinees in this sample scored slightly lower on Form C of Passage Comprehension than on Forms A and B. In the same way, users might conclude that the grades 3 through 4 and grades 9 through 12 examinees in this study scored slightly higher on Form A of Applied Problems than on Forms B and C. Additionally, all three samples of examinees in this study appeared to score slightly higher on Form C of Test 3: Spelling than on Forms A and B. Further research is warranted to determine whether these findings are specific to these particular samples of examinees (suggesting specialized knowledge or curricular exposure within these samples).

Table 6.
Percentage of Examinees
With Overlapping
W-Score 68%
Confidence Bands for
the WJ IV ACH Standard
Nonspeeded Tests

Study Group	Form Comparison	Study 1A: Reading		Study 1B: Mathematics		Study 1C: Written Language	
		Test 1: Letter-Word Identification	Test 4: Passage Comprehension	Test 2: Applied Problems	Test 5: Calculation	Test 3: Spelling	Test 6: Writing Samples
Grades 3–4	A & B	78.0%	73.2%	65.0%	87.5%	82.5%	75.0%
	A & C	87.8%	70.7%	77.5%	82.5%	82.5%	—
	B & C	82.9%	80.5%	87.5%	85.0%	67.5%	85.0%
Grades 9–12	A & B	75.0%	70.0%	60.0%	82.5%	80.0%	65.0%
	A & C	82.5%	80.0%	80.0%	77.5%	60.0%	75.0%
	B & C	77.0%	75.0%	90.0%	82.5%	62.5%	—
College+	A & B	90.0%	85.0%	82.1%	94.9%	85.0%	—
	A & C	80.0%	62.5%	82.1%	82.1%	72.5%	70.0%
	B & C	80.0%	60.0%	89.7%	79.5%	77.5%	90.0%

In contrast, an examination of Table 6 also reveals that there are several form pairs that have relatively high percentages of examinees with overlapping 68% confidence bands, despite those form pairs having significantly different mean scores in these samples (see Tables 4 and 6). For instance, Tables 4 and 5 show that the mean examinee W score for grades 3 through 4 on Form A of Test 5: Calculation was significantly lower than the mean examinee W scores on Forms B and C. However, the percentage of examinees with overlapping 68% score confidence bands for this study group in Table 6 was 87.5% for Forms A and B and 82.5% for Forms A and C, suggesting that the observed score differences might be due more to random error in the observed measures than to actual form difficulty differences. Again, further studies of form equivalence employing samples with varying demographic characteristics and across multiple settings are warranted.

To determine how consistently the three forms of the WJ IV ACH Standard nonspeeded tests rank order examinees in these samples, Pearson correlations were computed for each test pair. Table 7 contains the sample sizes and coefficients of equivalence for each test form pair in each study group. During the analysis of this data, it was noted that the variability of examinee W scores was much lower for these samples when compared to the same-age sample from the WJ IV norming study. This range restriction might be due to the homogeneity of these study samples with respect to race,

ethnicity, and socio-economic status/educational attainment. To account for attenuation in the correlations due to range restriction, a correction (see Alexander, 1990; Bryant & Gokhale, 1972; and Sackett & Yang, 2000) was applied to correlations when the ratio of the alternate-forms study sample *W*-score standard deviation (*SD*) to the corresponding WJ IV norming sample *W*-score *SD* was .80 or less.

Table 7.
*Correlations of
Equivalence for the
WJ IV ACH Standard
Nonspeeded Tests*

Study Group	Form Comparison	Study 1A: Reading					
		Test 1: Letter-Word Identification			Test 4: Passage Comprehension		
		<i>n</i>	<i>r</i> Original	<i>r</i> Corrected	<i>n</i>	<i>r</i> Original	<i>r</i> Corrected
Grades 3–4	A & B	41	.88	.95	41	.74	.96
	A & C	41	.92	.97	41	.64	.94
	B & C	41	.93	.96	41	.82	.97
Grades 9–12	A & B	40	.77	.92	40	.64	.86
	A & C	40	.79	.93	40	.78	.91
	B & C	40	.79	.93	40	.64	.87
College+	A & B	40	.78	.93	40	.74	.91
	A & C	40	.75	.92	40	.80	.93
	B & C	40	.73	.94	40	.77	.92

Study Group	Form Comparison	Study 1B: Mathematics					
		Test 2: Applied Problems			Test 5: Calculation		
		<i>n</i>	<i>r</i> Original	<i>r</i> Corrected	<i>n</i>	<i>r</i> Original	<i>r</i> Corrected
Grades 3–4	A & B	40	.80	.90	40	.87	.95
	A & C	40	.82	.91	40	.89	.97
	B & C	40	.85	.92	40	.86	.95
Grades 9–12	A & B	40	.85	.91	40	.88	.94
	A & C	40	.84	.91	40	.83	.92
	B & C	40	.88	.93	40	.84	.92
College+	A & B	39	.88	—	39	.92	.95
	A & C	39	.86	—	39	.90	.94
	B & C	39	.92	—	39	.88	.95

Study Group	Form Comparison	Study 1C: Written Language					
		Test 3: Spelling			Test 6: Writing Samples		
		<i>n</i>	<i>r</i> Original	<i>r</i> Corrected	<i>n</i>	<i>r</i> Original	<i>r</i> Corrected
Grades 3–4	A & B	40	.91	.96	20	.12	.87
	A & C	40	.93	.97	—	—	—
	B & C	40	.91	.96	20	.08	.94
Grades 9–12	A & B	40	.81	.94	20	.34	.86
	A & C	40	.75	.91	20	.81	.94
	B & C	40	.82	.94	—	—	—
College+	A & B	40	.83	.94	—	—	—
	A & C	40	.79	.93	20	.80	.93
	B & C	40	.84	.94	20	.74	.90

Note. In cases where the ratio of the *W*-score standard deviation in the alternate-forms study sample to the *W*-score standard deviation in the norming sample was .80 or less, a correction for attenuation due to range restriction was applied to the correlation. The correction was originally developed by Bryant and Gokhale (1972) and Alexander (1990) and is described by Sackett and Yang (2000). It is particularly suited for cases where unrestricted variances for both correlated variables are known.

Corrected correlations in Table 7 are all strong, ranging from .86 to .97. Users are cautioned about interpreting the corrected correlations in Table 7 due to the severe restriction in range of W scores for several of the age-group samples in this study. Across all examinee age-group samples and tests, the ratios of alternate-forms study sample W-score SDs to norming sample W-score SDs ranged from .36 to .88 (median = .60). Clearly, this sample was much more heterogeneous than the WJ IV norming sample, which limits the generalizability of the correlations between forms. The restriction of examinee W-score ranges in this study was most exaggerated for Test 6: Writing Samples, with sample SD-to-norming SD ratios ranging from .36 to .62 across forms and study groups. These low ratios necessitated a large correction, which may have resulted in inflated coefficients of equivalence for Test 6: Writing Samples. Additionally, because of the study sampling design, the Test 6: Writing Samples correlations are based on examinee samples that are roughly half the size of the other test samples, limiting the power of the correlational analysis for that test.

Speeded Tests Alternate-Forms Equivalence Studies

A second study was conducted to determine the extent to which the average item difficulties across the three forms of each of the WJ IV ACH Standard speeded tests are equivalent. If the three forms of a speeded test are equivalent—in other words, if the average item difficulty on each form is the same—then the examinee rates of correct response should be approximately equal across forms.

Data for this study were gathered from March 2011 through January 2012. Examiners who had been hired and trained by Riverside recruited convenience samples of study participants in the following three groups: grades 3 through 4, grades 9 through 12, and adults with at least a high school education. Examinees could not have participated in a prior Woodcock-Johnson IV research study. Individuals with a diagnosis of cognitive or developmental delay, and individuals with IEPs or 504 plans, were not included in the studies. After each participant (or their parent or guardian) completed a consent form, the examiner randomly assigned the individual to one of six counterbalancing conditions, which varied the order of administration for Test 9: Sentence Reading Fluency, Test 10: Math Facts Fluency, and Test 11: Sentence Writing Fluency. Study participants were administered all three forms of Sentence Reading Fluency and Math Facts Fluency but only two forms of Sentence Writing Fluency. This design allowed for sufficient counterbalancing to ensure that any significant differences between forms were not due to fatigue or practice effects, while also minimizing the chance of examinee fatigue on the sentence writing task. Table 8 contains the total number and distribution of examinees in each administration condition for the speeded alternate forms study.

Table 8.
*Distribution of Examinees
in the WJ IV ACH Standard
Speeded Tests Alternate
Forms Study Group by Age
and Administration Order*

	Test Administration Order	Grades 3–4	Grades 9–12	Adult	Total
Test 9: Sentence Reading Fluency	A-B-C	40	44	50	134
	B-C-A	40	46	48	134
	C-A-B	40	43	51	134
	Total	120	133	149	402
Test 10: Math Facts Fluency	A-B-C	40	44	51	135
	B-C-A	40	45	42	127
	C-A-B	40	42	53	135
	Total	120	131	146	397

Table 8. (cont.)
*Distribution of Examinees
in the WJ IV ACH Standard
Speeded Tests Alternate
Forms Study Group by Age
and Administration Order*

	Test Administration Order	Grades 3–4	Grades 9–12	Adult	Total
Test 11: Sentence Writing Fluency	A-B	19	21	26	66
	B-A	20	18	26	64
	B-C	19	24	21	64
	C-B	20	22	26	68
	C-A	19	22	28	69
	A-C	20	24	26	70
	Total	117	131	153	401

Table 9 contains demographic information about the samples for the WJ IV ACH Standard speeded tests alternate forms study. Because study participant groups were samples of convenience, demographic characteristics are not necessarily representative of the U.S. population. Within each study group, the samples of examinees predominantly are White and Not Hispanic. For the grades 3 through 4 and grades 9 through 12 study groups, the level of parent education generally is higher than in the U.S. population. For the adult sample, the level of examinee educational attainment also is significantly higher than the level of educational attainment for adults of the same age in the U.S. population.

Table 9.
*Sample Demographic
Characteristics for the WJ IV
ACH Standard Speeded Tests
Alternate-Forms Study*

Demographic Characteristics	Study Group		
	Grades 3–4	Grades 9–12	Adults
	<i>N</i> = 120	<i>N</i> = 142	<i>N</i> = 166
Age (years)			
Range	7–10	13–19	18–73
Mean	9.1	15.9	36.0
<i>SD</i>	0.7	1.3	12.0
Sex			
Male	45.0	49.3	40.4
Female	55.0	50.7	59.3
Race			
White	88.3	78.9	89.8
Black	9.2	16.9	6.6
American Indian/Alaska Native	0.8	0.7	—
Asian/Native Hawaiian/Pacific Islander	1.7	2.1	1.8
Other/Mixed	—	1.4	1.8
Ethnicity			
Not Hispanic	96.7	93.6	97.6
Hispanic	3.3	6.3	2.4
Mother's Education			
No Information Provided		0.7	—
< High School Graduate	10.8	7.7	—
High School Graduate	11.7	23.9	—
> High School	77.5	67.6	—

Table 9. (cont.)
Sample Demographic
Characteristics for the WJ IV
ACH Standard Speeded Tests
Alternate-Forms Study

Demographic Characteristics	Study Group		
	Grades 3–4	Grades 9–12	Adults
	N = 120	N = 142	N = 166
Father's Education			
No Information Provided	8.3	7.0	—
< High School Graduate	7.5	5.6	—
High School Graduate	25.0	21.8	—
> High School	59.2	65.5	—
Examinee's Grade in School			
Grade 3	45.8		—
Grade 4	54.2		—
Grade 9		47.9	—
Grade 10		14.1	—
Grade 11		17.6	—
Grade 12		20.4	—
Examinee's Highest Year of Education Completed			
Grade 12	—	—	4.4
First Year of College	—	—	20.0
Second Year of College	—	—	14.5
Third Year of College	—	—	3.6
Bachelor's Degree	—	—	44.2
Master's Degree or Higher	—	—	13.3

Table 10 contains study group means and SDs for each form of the WJ IV ACH Standard speeded tests. For ease of interpretation, the means and SDs in Table 10 are presented in a rate-based metric, which is *related* to (but is not *equal* to) the examinee rate of correct response per minute of testing.³ Mean examinee scores and standard deviations were consistent across all tests and study groups, suggesting that the difficulties of the items on each form are, on average, approximately equal. Although there may be slight differences in item difficulties from form to form, when raw item scores are converted into a rate-based metric, the results in Table 10 suggest that any slight differences in item difficulty seem to cancel out, resulting in an overall rate of correct response that is very consistent from form to form. Mean scores on Test 11: Sentence Writing Fluency show larger between-form differences than the other tests do. However, because each examinee took only two of the three forms of Test 11: Sentence Writing Fluency, the summary statistics reported in Table 10 for that test are not from completely matched samples. In other words, the slight differences in examinee mean scores on Test 11: Sentence Writing Fluency may be due to differences in the abilities of the examinees who took each form rather than differences in item difficulty between forms.

³ During the calibration of the WJ IV speeded tests, scaling constants were introduced to account for the length of each test and to be consistent with the data collected from administration of the tests during the norming study. Although the rate-based scores reported in Table 10 can be described as rates of correct response within each minute of testing, they should not be strictly interpreted as “the number of correct responses per minute.” In the Woodcock-Johnson online scoring and reporting program (Schrack & Dailey, 2014), the calculation of the rate-based score and the subsequent conversion to a W score is completed by the software; the rate-based data are provided in Table 10 simply to assist the user in interpreting the comparison of examinee performance across forms.

Table 10.
Mean Examinee Rate-
Based Scores and Standard
Deviations for the WJ IV
ACH Standard Speeded
Tests

Study Group	Form	Test 9: Sentence Reading Fluency		Test 10: Math Facts Fluency		Test 11: Sentence Writing Fluency	
		Mean	SD	Mean	SD	Mean	SD
Grades 3–4	A	7.2	2.0	9.2	3.4	16.0	4.4
	B	7.3	2.0	9.0	3.4	17.0	4.6
	C	6.9	2.0	9.3	3.4	16.5	4.8
Grades 9–12	A	12.2	2.7	17.3	4.0	28.2	7.2
	B	12.5	2.7	17.4	4.1	28.2	5.6
	C	12.0	2.5	17.5	4.2	30.7	6.3
College+	A	14.3	3.0	22.1	5.5	30.2	5.8
	B	14.4	3.3	22.0	5.5	30.3	5.2
	C	14.1	3.0	21.9	5.5	32.1	5.2

Table 11 contains Pearson correlations between each set of test forms for each study group. Correlations between alternate forms of Test 9: Sentence Reading Fluency are moderately strong, ranging from .82 to .88. Correlations for Test 10: Math Facts Fluency are strong, ranging from .92 to .96. Correlations for Test 11: Sentence Writing Fluency show more variability, ranging from .76 to .91. These patterns of correlations are not unexpected when considering the test content and the study sampling design. For example, Test 10: Math Facts Fluency has the highest between-form correlations and is the test with the greatest match in individual item content across the three forms—items that are similar across forms (e.g., $2 + 1$ versus $1 + 2$) and items that are exact matches but in slightly different locations across forms. The relatively large range of correlations for Test 11: Sentence Writing Fluency might be due to the fact that each examinee took only two of the three forms of this test; thus, each correlation is based on only approximately one third of the examinee scores in each sample. Overall, the correlations presented in Table 11 show that the three forms of the speeded tests similarly rank-ordered the examinees in this sample. This provides further support for the equivalence of the alternate forms of the WJ IV ACH Standard speeded tests.

Table 11.
Correlations of Equivalence
for the WJ IV ACH Standard
Speeded Tests

Study Group	Form Comparison	Test 9: Sentence Reading Fluency		Test 10: Math Facts Fluency		Test 11: Sentence Writing Fluency	
		<i>n</i>	Correlation	<i>n</i>	Correlation	<i>n</i>	Correlation
Grades 3–4	r_{AB}	120	0.87	120	0.95	39	0.84
	r_{AC}	120	0.87	120	0.94	39	0.84
	r_{BC}	120	0.88	120	0.95	39	0.91
Grades 9–12	r_{AB}	133	0.85	131	0.92	39	0.82
	r_{AC}	133	0.88	131	0.93	46	0.89
	r_{BC}	133	0.85	131	0.94	46	0.76
College+	r_{AB}	149	0.82	146	0.96	52	0.87
	r_{AC}	149	0.87	146	0.96	47	0.91
	r_{BC}	149	0.88	146	0.95	54	0.79

Summary

The WJ IV ACH Standard is the only individually administered achievement test available in three forms, making it well suited for use in settings where different forms of the tests are used for different purposes or when an individual may be tested more than once after a reasonable period of instruction (such as after 6 to 9 months of an intensive program of interventions or special services). The purpose of this bulletin is to provide professionals with information about the equivalence of the WJ IV ACH Standard alternate forms. Care was taken during the item calibration, equating, scaling, and form assembly stages of test development to engineer three alternate forms of each test that are as equivalent as possible with regard to content coverage, overall item difficulty, and score precision.

Using the publication version Forms A, B, and C of each test, data were gathered from several hundred examinees across three study groups. The equivalence of the forms was evaluated through examination of mean-score differences, percentage of examinees with overlapping *W*-score confidence bands (for the nonspeeded tests), and correlations of equivalence for each pair of forms. Overall, the evidence supports the equivalence of the three forms of the tests for the examinee ages included in these studies.

However, as in any study where measurement error or error due to study design or sampling are considerations, some differences were observed. Any apparent score differences were evaluated in the context of both statistical significance and practical meaning. A method for assessing the meaningfulness of apparent score differences using the *SEM* associated with each examinee's *W* score was employed. Correlations of equivalence were moderate to strong across all study groups for both the nonspeeded and speeded tests. Limitations of these studies, including samples that are homogeneous with regard to race, ethnicity, and socio-economic status/educational attainment, were discussed. Table 12 contains a summary of the evidence for the equivalence of the alternate forms of the WJ IV ACH Standard tests.

Table 12.
*Summary of Evidence for
Equivalence of WJ IV ACH
Standard Alternate Forms*

Criteria for Evaluation of Form Equivalence	Evidence Presented From		
	Documentation of Test Construction	Nonspeeded Tests Alternate Forms Study	Speeded Tests Alternate Forms Study
Equivalence of content coverage	Content-area expert reviews ensured comparability of item types.	—	—
Equivalence of form difficulty	Figures 1–16 demonstrate excellent equivalence of item difficulty and raw score-to- <i>W</i> score conversions.	Tables 4–6 demonstrate very good equivalence for the Reading and Mathematics tests and acceptable equivalence for the Written Language tests.	Table 10 demonstrates excellent equivalence for the Reading and Mathematics tests and very good equivalence for the Written Language tests.
Equivalence of measurement precision	Figures 17–24 demonstrate very good equivalence of <i>SEM</i> across <i>W</i> -score ranges.	—	—
Equivalence of rank-ordering of examinee ability	—	Table 7 demonstrates high to very high correlations of equivalence for all tests.	Table 10 demonstrates high to very high correlations of equivalence for all tests.

In summary, professionals should feel confident that the three forms of the WJ IV ACH Standard tests are equivalent for their intended purposes. Professionals should always apply knowledge of measurement error as well as any individual patterns of learning when comparing scores across forms for any given individual.

References

- Alexander, R. A. (1990). Correction formulas for correlations restricted by selection on an unmeasured variable. *Journal of Educational Measurement*, 27, 187–189.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Bryant, N. D., & Gokhale, S. (1972). Correcting correlations for restrictions in range due to selection on an unmeasured variable. *Educational and Psychological Measurement*, 32, 305–310.
- Hinton-Bayre, A., & Geffen, G. (2005). Comparability, reliability, and practice effects on alternate forms of the Digit Symbol Substitution and Symbol Digit Modalities tests. *Psychological Assessment*, 17, 237–241.
- McCallum, R. S., & Bracken, B. A. (1981). Alternate form reliability of the PPVT-R for white and black preschool children. *Psychology in the Schools*, 18, 422–425.
- McCarty, S. M., Ziesat, H. A., Logue, P. E., Power, D. G., & Rosenstiel, A. K. (1980). Alternate-form reliability and age-related scores for Russell's Revised Wechsler Memory Scale. *Journal of Consulting and Clinical Psychology*, 48, 296–298.
- McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical Manual. *Woodcock-Johnson IV*. Rolling Meadows, IL: Riverside.
- Ross, T. P., Furr, A. E., Carter, S. E., & Weinberg, M. (2006). The psychometric equivalence of two alternate forms of the controlled oral word association test. *The Clinical Neuropsychologist*, 20, 414–431.
- Ryan, J. J., Geisser, M. E., Randall, D. M., & Georgemiller, R. J. (1986). Alternate form reliability and equivalence of the Rey Auditory Verbal Learning Test. *Journal of Clinical and Experimental Neuropsychology*, 8, 611–616.
- Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85, 112–118.
- Schmidt, K. S., Mattis, P. J., Adams, J., & Nestor, P. (2005). Alternate-form reliability of the Dementia Rating Scale-2. *Archives of Clinical Neuropsychology*, 20, 435–441.
- Schrank, F. A., & Dailey, D. (2014). *Woodcock-Johnson online scoring and reporting* [Online format]. Rolling Meadows, IL: Riverside.
- Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV Tests of Achievement*. Rolling Meadows, IL: Riverside.
- U.S. Census Bureau. (2014). *QuickFacts*. Retrieved from <http://quickfacts.census.gov/qfd/states/00000.html>
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.



**Houghton
Mifflin
Harcourt™**

3800 Golf Road, Suite 200
Rolling Meadows, IL 60008

800.323.9540
www.wj-iv.com

Riverside