Equivalence of Remote and In-Person Administration of the Woodcock-Johnson V Tests of Cognitive Abilities and Achievement

Stephanie Taylor, Ed.S., NCSP
Taylored Education Solutions

July, 2025

Abstract

This study evaluated the equivalency of remote and in-person administration of the Woodcock-Johnson V (WJ V) Tests of Cognitive Abilities and Achievement, using a matched case-control design with 300 participants aged 5–18. Participants were randomly assigned to remote or in-person administration and matched by age, gender, and ethnicity. Remote assessments were conducted using the Presence Platform under controlled conditions—including 13" laptops, high-resolution dual-camera setups, directed proctors, embedded content from the publisher, a specific set of test security protocols, and licensed examiners trained in both WJ V administration and the digital platform. Results demonstrated no statistically or practically significant differences between administration methods ($p > .05$; $\omega^2 < .03$). These findings confirm the equivalency of remote administration under these strict fidelity standards.

Introduction

Traditional, in-person administration has long served as the foundation of psychoeducational assessment. However, advancements in digital tools and growing needs across school systems have prompted increased interest in remote assessment options. These options aim to address nationwide staffing shortages, spikes in referral volumes, logistical challenges in securing staffing and the rise of alternate learning environments such as virtual schools, while also freeing up onsite staff to offer proactive services and behavioral supports to traditional settings.

Companies like Presence, which served as the platform for this study, have been providing remote psychoeducational assessments since 2016. Since COVID, there has been a rise in the use of remote school psychology services and assessments. This has led to a need for more research around which assessments demonstrate equivalency and under what conditions is equivalency demonstrated. Those conditions must pave the way for a standard assessment protocol. Clear guidance remains a critical need in the field, as many practitioners are approaching remote assessment with varying experience and level of guidance. While some research is available around the remote administration of major assessments, more is needed, including advocacy around the specific conditions that must be met for remote assessment to be considered equivalent.

This study investigates the equivalence of remote and in-person administration formats for the Woodcock-Johnson V (WJ V) Tests of Cognitive Abilities and Tests of Achievement, published by Riverside Insights. The primary objective is to determine whether scores obtained through remote administration using the Presence Platform are statistically and practically equivalent to those obtained through traditional, in-person methods. These results contribute to the expanding body of research around determining a standard protocol to inform school psychologists, educators, and decision-makers about the viability of remote administration and the conditions under which it must be conducted.

Multiple factors impact the validity and reliability of scores. As with traditional, in-person assessment, consideration must be given to the student environment, test security, and examiner experience. Additionally, remote administration requires considering the original format of the assessment being given (digital vs. paper), the telecommunications platform, and the equipment being used.

Fidelity Requirements

Remote administration of the WJ V was conducted under tightly controlled conditions to ensure fidelity, outlined below. It's important to note that the results of this study should only be applied to remote testing scenarios that follow the specific fidelity standards used in the research. Other forms of remote or online test administration have not been examined for their equivalence to traditional in-person testing.

Physical Conditions

Remote administration of the WJ V must occur in a quiet, distraction-free room with the student seated in front of a touchscreen device. While the WJ V was normed using iPads, iPads are often not conducive to remote assessment for multiple reasons. For this study, a touchscreen laptop of 13 inches or larger was used for the student device. The student interacts with the examiner both verbally and visually through a secure digital platform. Two high-definition cameras are used—one focused on the student's face and a document camera showing their workspace. The student can see the examiner and the digital test materials on their screen. A proctor remains in the room to support the process by handling materials, providing technical support as needed, and following the examiner's instructions for examinee interactions.

Digital Platform

The Presence Platform was used exclusively for remote testing. As the WJ V is already a digital assessment with a fully digital interface, it was simply embedded into the Presence Platform to give the user full access to the stimulus and user interface. The Presence Platform provided the access to the multiple camera feeds, clear audio and video, and HIPAA compliant security. Stimulus materials were presented digitally on the platform through Riverside Score, while student response booklets remained physical. In order to maintain the fidelity of the stimuli and allow space for the camera feeds, a 13" screen was required. Touchscreen devices were required to maintain compliance with assessment norms.

For this study, students were given a URL to join, which allowed them to "knock" on the examiner's door to their virtual testing room. Having the URL did not allow them to enter before the examiner. The Proctor then completed a "tech check" to assure their equipment met the audio, video, and operating needs of the session.

Examiner Training

All of the study examiners were licensed school psychologists with some previous knowledge of the Woodcock Johnson, primarily the IV. Most examiners had not yet switched to using the WJ V and had little experience with it. Each examiner completed a one-hour virtual training session led by the study director to orient them to the changes from the WJ IV to the WJ V, and to familiarize themselves with the Presence Platform. The platform training consisted of a live demonstration and access to ongoing reference materials. They were also given voluntary access to ongoing office hours and supports as needed.

Proctor Training

Proctors were recruited by individual examiners and were adults over the age of 18. They were trained individually by examiners using a 3-page PDF of the proctor duties, provided by the study director. The PDF included how to complete the “tech check,” how to enter the testing environment, how to verify audio and video were working, and explicit instructions to refrain from interacting with the examinee or the platform unless directed by the school psychologist.

Study Design

This study utilized a case-control match design format. Examinees were paired by age, gender, and ethnicity and were randomly assigned be to either the in-person or the remote group. The two samples (in-person and remote) were equal in number and should generally equivalent on potentially confounding variables and general cognitive ability (because of the random assignment). For the purposes of this study, both two-tailed t-tests (resulting in p-values) and effect sizes (\omega^2) were calculated to assess the equivalency of the administration methods. The standards of $p \geq .05$ and $\omega 2 \leq .03$ were selected as cut-off criteria for a significant effect of the administration type. Omega squared ($\omega 2$) was selected based on the design and power with the sample size, and the criterion cut-off was selected to correct for power issues with a small sample size (see Rice, 2006).

Participants

The sample includes 300 participants between the ages of 5 and 18. Participants were matched by age, gender, and ethnicity. Examinees were recruited by 44 trained examiners from diverse regions across the United States. Each examiner contributed participants to both the in-person and remote conditions.

Table 1. Sample size summary by administration method, student gender, age, ethnicity, parent education.

		Administration Method			
		Online		In-person	
Demographic Characteristic	Category	Male	Female	Male	Female
Age Group (years)	5-6	8	8	8	8
	7-8	8	8	8	8
	9-10	12	12	12	12
	11-12	12	12	12	12
	13-14	12	12	12	12
	15-16	12	12	12	12
	17-18	11	11	11	11
Ethnicity	Asian	5	5	5	5
	Black	10	10	10	10
	Latino	19	19	19	19
	Native American	2	2	2	2
	White	39	39	39	39
Parent's Education	1st year of college+	56	61	60	64
	High school graduate	10	8	7	8
	K-11	9	6	8	3
Total		150		150	

Procedures

Data were collected between March and June 2025 using standardized administration procedures across both formats. 44 trained examiners participated in the study, each conducting both in-person and remote assessments. Outside of the experimental conditions, all administrations adhered strictly to the standardized testing protocols as outlined in the WJ V manuals.

Both the in-person and remote groups utilized touchscreen laptops for examinees. Touchscreen devices are essential for several of the timed subtests. The norming tables are based on touchscreen device use. The use of a mouse or trackpad would invalidate the timed subtests that utilize the touchscreen such as Number-Pattern Matching, Sentence Reading Fluency, and others. For this study, a touchscreen laptop was the ideal device, allowing enough screen space for all of the camera feeds and the stimulus, placing the camera in the optimal position to see the student, and connecting easily to a document camera.

Results

The means and standard deviations of the WJ V subtest standard scores for each administration format and for the total sample are presented in Table 2.

Table 2: Means and standard deviations of scores by test and administration method compared to overall means and standard deviations.

Mean and SD of Test Scores by Administration Method

	In-Person		Online		Overall	
	Mean	SD	Mean	SD	Mean	SD
Letter-Word Identification	101.61	12.80	101.69	12.55	101.65	12.65
Calculation	101.03	13.53	101.21	15.14	101.12	14.33
Spelling	99.67	13.72	100.58	13.62	100.13	13.66
Math Facts Fluency	98.66	12.67	97.27	14.30	97.96	13.51
Passage Comprehension	102.37	13.29	103.45	16.30	102.91	14.85
Applied Problems	102.98	15.28	102.37	16.40	102.68	15.83
Sentence Reading Fluency	94.16	12.69	94.71	12.81	94.43	12.73
Written Language Samples	105.71	14.71	104.08	15.61	104.89	15.17
Sentence Writing Fluency	99.02	16.00	96.73	18.96	97.88	17.55
Broad Achievement	100.81	13.96	101.16	14.14	100.98	14.03
Oral Vocabulary	105.20	15.58	103.63	15.69	104.42	15.63
Matrices	106.46	15.18	104.69	15.28	105.57	15.23
Spatial Relations	105.67	13.01	102.64	14.47	104.15	13.82
Story Recall	105.89	14.78	103.28	16.18	104.58	15.53
Semantic Retrieval	108.28	14.43	106.09	15.75	107.18	15.12
Verbal Analogies	105.01	15.52	104.45	17.68	104.73	16.61
Number-Pattern Matching	94.93	12.38	94.45	12.98	94.69	12.66
Verbal Attention	102.33	14.33	100.86	14.88	101.59	14.60
GIA	103.18	13.57	101.63	15.04	102.40	14.32

Note: All scores are standard scores (M= 100, SD = 15).

Data was analyzed using two types of statistical tests to assess the equivalence of the administration methods. First a two-tailed t-tests was conducted to test for a difference in means with an a priori type I error rate of = 0.05; if the p-value resulting from the t-tests is greater than 0.05 it can be concluded there is no statistical difference. Additionally, a one-way analysis of variance (ANOVA) on each of the test categories and sub categories was conducted to estimate the amount of total variation in test scores that can be attributed to the administration method (commonly referred to as effect size). An a priori cutoff of 0.03 was utilized for effect size; if the estimated effect size is less than 0.03, it can be concluded there is no statistical difference. The P-value column in Table 3 indicates the statistical significance of the t-tests, and the 2 column indicates the estimated effect size from the ANOVA.

Table 3. Statistical test results for two-sided t-test and ANOVA effect size. The estimated difference is the average in person score minus the average remote score.

T-Test and Effect Size Results by Test

Test	Estimated Difference	T-Statistic	P-Value	ω^2
Letter-Word Identification	-0.073	-0.050	0.960	0.000
Calculation	-0.173	-0.105	0.917	0.000
Spelling	-0.907	-0.574	0.566	0.000
Math Facts Fluency	1.393	0.893	0.373	0.000
Passage Comprehension	-1.080	-0.629	0.530	0.000
Applied Problems	0.607	0.331	0.741	0.000
Sentence Reading Fluency	-0.547	-0.371	0.711	0.000
Written Language Samples	1.627	0.929	0.354	0.000
Sentence Writing Fluency	2.287	1.129	0.260	0.001
Broad Achievement	-0.353	-0.218	0.828	0.000
Oral Vocabulary	1.567	0.868	0.386	0.000
Matrices	1.773	1.008	0.314	0.000
Spatial Relations	3.027	1.905	0.058	0.009
Story Recall	2.607	1.456	0.146	0.004
Semantic Retrieval	2.193	1.258	0.209	0.002
Verbal Analogies	0.553	0.288	0.774	0.000
Number-Pattern Matching	0.480	0.328	0.743	0.000
Verbal Attention	1.467	0.870	0.385	0.000
GIA	1.553	0.939	0.348	0.000

P-values are greater than 0.05 and all the estimated effect sizes are less than 0.03 indicating the in-person and remote administration methods are statistically equivalent.

Discussion

This study aimed to determine equivalency of in-person and remote administration of the Woodcock Johnson V Tests of Cognitive Abilities and Tests of Achievement.

Analysis shows remote administration of the WJ V tests to be equivalent to in-person administration. The statistical analyses of the test scores indicated high agreement between the administration methods when the equipment, environment, training, and proctor conditions are met.

However, it is noteworthy that the Spatial Relations subtest and, to a lesser extent, Story Recall, indicated a nearly statistically significant difference between methods with p-values of 0.058 and 0.146 respectively. These near significant results should be interpreted with a caveat however, as when conducting multiple hypothesis tests (more than 2), it is common practice to decrease the a priori type I error rate, to less than the nominal 0.05, or upward adjust the p-values to account for the increased probability of discovering an anomalous, statistically significant difference. If the Bonferroni multiple comparisons adjustment was used, the cutoff for determining a difference in methods would change from =0.05 to =0.0026, which all of the p-values in table 3 are significantly greater than (see Rice, 2006).

References

Castillo, J. M., Curtis, M. J., & Tan, S. Y. (2014). Personnel needs in school psychology: A 10-year follow-up study on predicted personnel shortages. *Psychology in the Schools, 51*(8), 832–849. https://doi.org/10.1002/pits.21786

Reynolds, M. R., & Niileksela, C. R. (2015). Test review: Schrank, F. A., McGrew, K. S., & Mather, N. (2014). Woodcock-Johnson IV Tests of Cognitive Abilities. *Journal of Psychoeducational Assessment, 33*(4), 381–390. https://doi.org/10.1177/0734282914563910

Rice, J. A. (2006). *Mathematical statistics and data analysis* (3rd ed.). Duxbury Press.

Schrank, F. A., McGrew, K. S., & Mather, N. (2025). *Woodcock-Johnson V Tests of Cognitive Abilities and Tests of Achievement: Examiner's manual*. Riverside Insights.

Taylor, S. (2023). *Equivalence of remote, online administration and traditional, face-to-face administration of psychoeducational assessments* [White paper]. Presence. https://www.presence.com

Woodcock, R. W., Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities and Tests of Achievement*. Riverside Insights.

Wright, A. J. (2018). *Essentials of tele-assessment*. Wiley. https://doi.org/10.1002/9781119485075