# Equivalence of Remote, Online Administration and Traditional, Face-to-Face Administration of *Woodcock-Johnson IV* Cognitive and Achievement Tests

A. Jordan Wright, PhD, ABAP
Empire State College, SUNY

September, 2016

## Introduction

Remote, online administration of cognitive and academic achievement tests has the potential to increase access to these services for many children and students. However, while the current state of technology offers excellent potential, it also poses methodological challenges for actual clinical work. PresenceLearning, Inc., in collaboration with Houghton Mifflin Harcourt, has developed a platform and process for the synchronous administration of the *Woodcock-Johnson IV*® (WJ IV™; Schrank, McGrew, & Mather, 2014) cognitive and achievement tests remotely-the examiner can be anywhere, at any distance from the examinee, as long as certain requirements are met.

With remote, online administration, the examinee sits in a secure location with a computer and a proctor overseeing and assisting in the process. The examiner sits at her or his own computer in a secure location. The stimulus materials are built into the online system, so that the examiner does not have to use traditional manuals or stimulus books. Two cameras allow the examiner to see the face and hands/workspace of the examinee, and the examinee can see the face of the examiner, in addition to stimulus materials, on the screen. Tests that require the Response Booklet are administered in the same way as the traditional, in-person administration, directly into the Response Booklet, which has a camera specifically trained on it so the examiner can see what is being written.

The present study evaluates the equivalence of the scatter of scores captured by remote, online administration of the WJ IV (Schrank et al., 2014) and traditional, in-person administration. The goal was to evaluate both cluster and individual test standard scores, to determine if the scores are interchangeable in the two formats, and if so that the current normative and psychometric information can be applied to both administration formats.

In theory, several factors could contribute to differences between traditional, in-person administration and online, remote administration of cognitive and achievement tests. From more obvious differences, such as experiencing and interacting with the stimuli differently in the two formats, to more subtle ones, such as influences of not having the examiner there in person, or of having a proctor overseeing the examinee for the duration of the test and interacting with both the examiner and the examinee, administering tests via an online platform introduces multiple potential factors that could affect examinee performance.

Specific steps were taken in the development of the online platform to minimize, as much as possible, the difference in experience between the online, remote administration and traditional, in-person administration. Some modifications to administration prompts, careful consideration of placement of stimulus material on the screen, and careful use of the proctor at specific moments in administration, for example, were utilized. Tasks with physical manipulatives, such as pencil and paper for performing mathematical calculations or spelling, were retained in their original format in the online, remote administration, rather than attempting to build them into the digital platform somehow.

PresenceLearning, Inc., in addition to building the platform, also developed a manual for adapting test administration to the online, remote format. The manual includes training materials for examiners and proctors, as well as specific necessary digital conditions and administration instructions.

## Fidelity Requirements

It must be noted that this study evaluated a very specific set of protocols for delivering remote, online administration of the *Woodcock-Johnson IV* cognitive and achievement tests. The generalizability of any findings should be limited to trained test administrators who follow the specific protocol examined. Below is a list of criteria necessary to reproduce the procedure faithfully.

### Digital Platform

The digital platform itself has several requirements that must be met, in order to maintain fidelity. First, the platform itself must have at least 800 x 600 pixels of resolution quality on both monitors (optimally, and for the present study, at least 1920 x 1080). The content must be digitized and placed within the platform (rather than materials held up to a digital camera or some other process), and the audio must run through the platform itself to ensure clarity and quality, rather than playing audio on the examiner's side and having a microphone pick it up for the examinee to hear (alternatively, the proctor can play the audio on the examinee's side). The video platform should mirror the images it projects, so that anything held up for the examinee is not seen backward. The examinee should have a screen of at least 15", a separate mouse (not a track pad or a built-in laptop pad), and a headset. The proctor should have a headset as well.

For the purposes of this study, the content of the tests was licensed directly from the publisher and was placed on a secure platform. The platform required both the examiner to log in to the platform with a password, and the examiner had to then "admit" the examinee into the actual test. That is, having the URL did not allow the examinee to log in and see the test without the examiner explicitly allowing it at that moment (for test security).

### Examiner Training

Once they had demonstrated competence with administering the WJ IV in the traditional, face-to-face format, test examiners underwent a specific training that included an overview of the mechanics of the digital platform to be used, a detailed guide on how to practice, and a practical exam during which they administered the tests to the Project Director, to show competence. Each received six hours of group training and two hours of individual training in administering the WJ IV via the prescribed remote, online procedure, and each conducted several practice administrations before the study began, as well as passed the practical demonstration/exam with the lead trainer/Project Director. While this level of training may not be feasible for all examiners under all conditions, specific, guided practice on the digital platform is necessary in order to ensure comfort with both the modality and any alterations in test administration that the modality requires.

Proctor Training

Proctors should be given specific instructions on what is expected of them, including when they are and are not allowed to speak to the examinee. For the purposes of this study, proctors were given a confidentiality agreement to sign, had a 30-minute preparatory lesson about what to expect and how to interact with the examinee and examiner, received a written guide about how to facilitate the setting up of the cameras on the examinee's side, and learned strict instructions (with examples) about specific things that would invalidate the testing. It is strongly recommended that proctor training be scripted and standardized, in order to ensure clarity and comprehensiveness.

Examiner Conditions

The necessary physical and mechanical conditions on the examiner's side include reliable, high speed internet connection (upload and download speeds of at least 500 kbps, though optimally broadband), a high definition camera, a headset with microphone, and access to the specific digital platform that has been created for the purpose of the modified administration procedure. In addition to these requirements, examiners must communicate with the proctor who will be in the room with the examinee, in order to clearly discuss expectations and set up the examinee-side cameras appropriately.

Examinee-Side Conditions

The physical and mechanical setup on the examinee's side is extremely important. The examinee should be working on a computer with no smaller than a 15-inch screen, with a headset with a microphone and a separate mouse (not a track pad). The examinee should have a reliable, high speed internet connection (upload and download speeds of at least 500 kbps, though optimally broadband). He or she should be seated in a space that has the computer screen and mouse close enough to comfortably respond to stimuli, but also has some space for writing on the tests that require it.

The examinee's space requires two cameras. The first one must be high definition and pointed at the student's face (this allows for connection to be developed, as well as facial behavioral observations to be observable by the examiner). The second must be a high quality document camera that is pointed at the examinee's writing workspace. This camera must be a high quality document camera in order for the quality to be good enough to clearly see and score written items (such as on the Spelling test).

## Equivalence Study Design

For the present study, a case control match design was utilized, in which examinees took the WJ IV tests (both cognitive and achievement) in only one format (traditional, in-person or remote, online). While a matched design requires a larger sample, it avoids testing effects of test-retest or alternate forms designs. That is, when an examinee answers either the same type of problem

or the exact problem more than once, learning may have occurred (of the problem solving strategy, or the content) that can alter the performance the second time. The case control match design creates two groups that are matched on specific variables/characteristics (in this case age and gender), each group receiving only one format of test administration, as would be the case in clinical practice. Therefore, two groups equal in number, matched on age and gender, were created, with examinees randomly assigned to one of the two groups.

For the purpose of this study, both significance tests (*p* values of *t*-tests) and Cohen's *d* were calculated to determine equivalence. The standards of $p \geq .05$ and $d < 0.2$ were used as the standard for equivalence. Cohen's *d* is calculated as the difference between mean scores on the two different administration formats, divided by the pooled standard deviation of scores.

## Participant Recruitment and Selection

The present study recruited and selected examinee participants from partner schools with demographic characteristics similar to a school general population. The present study did not screen for any specific disorders (intellectual, learning, or otherwise) in order to maintain generalizability to the general population. No students with specific hearing, seeing, or physical impairments were utilized in the present study, however.

Examiners were recruited through participating school districts and through state psychological associations. They received thorough training (as described above in the Fidelity Requirements section) in order to facilitate the smooth administration of the tests in the digital format.

## WJ IV Equivalence Study

### Method

Case control matching (on age and gender) was used for the WJ IV equivalence study. Data were collected between January and May 2016 in California, Idaho, and Florida.

### Participants

The sample consisted of 240 children, ages 5 to 16, who were recruited by PresenceLearning, Inc. Participants were randomly assigned to either remote, online administration or traditional, in-person administration of the WJ IV, with equal numbers in each cell by age and gender (see Table 1). All examinees were paid for their participation in the study.

Table 1 reports demographic characteristics of the sample. The subgroups taking the WJ IV with standard or remote administration were very similar. Overall, there was equal representation of males and females. Latino and Native American children were slightly overrepresented compared with the general population, while Black, White, and Asian children were slightly underrepresented compared with the general population.

Table 1
Demographic Characteristics of the WJ IV Sample

| Demographic Characteristic | | Administration Format | | | |
|---|---|---|---|---|---|
| | | Traditional, In-Person | | Remote, Online | |
| | | Number of Cases | | | |
| | | 120 | | 120 | |
| | | Male | Female | Male | Female |
| Age (years) | 5-6 | 10 | 10 | 10 | 10 |
| | 7-8 | 10 | 10 | 10 | 10 |
| | 9-10 | 10 | 10 | 10 | 10 |
| | 11-12 | 10 | 10 | 10 | 10 |
| | 13-14 | 10 | 10 | 10 | 10 |
| | 15-16 | 10 | 10 | 10 | 10 |
| | Mean | 10.52 | | 10.57 | |
| | *SD* | 3.483 | | 3.439 | |
| Race/Ethnicity | Asian | 4.1% | | 3.2% | |
| | Black | 15.8% | | 4.1% | |
| | Latino | 31% | | 20.5% | |
| | Native American | .8% | | 4.1% | |
| | White | 48.3% | | 68% | |
| Parent Education | Less than HS graduate | 5 | | 7 | |
| | HS graduate | 16 | | 12 | |
| | College Experience | 99 | | 101 | |

Examiners were school and clinical psychologists qualified and experienced in administering the WJ IV. After demonstrating competence in administering the WJ IV in the traditional, face-to-face format, all examiners received six hours of group training and two hours of individual training in administering the WJ IV via the prescribed remote, online procedure, and they conducted several practice administrations before the study began, as well as passed a practical demonstration/exam with the lead trainer. All examiners were paid for their training and participation in the study.

Proctors were recruited from different sources. Some were volunteers who wanted to learn more about the WJ IV, some were students who were getting contact hours with "clients," and some were paid professionals (school psychologists). All proctors received 30 minutes of training directly before their first proctoring of the remote, online WJ IV.

## Procedure

As each participant was scheduled for testing, he/she was randomly assigned to either the traditional, in-person or the remote, online administration format, with the requirement that the cases within each age-by-gender "cell" would be divided equally between the formats. All administrations (in both formats) occurred within the child's school, and all examiners administered cases using both formats.

Each examinee first took the nonverbal tests of the *Cognitive Abilities Test*™ (CogAT®), Form 6 (Lohman & Hagen, 2001), which yields a nonverbal ability score. Every examinee was administered the CogAT in the traditional, paper-and-pencil format. Then each examinee took the complete WJ IV, both cognitive and achievement tests, in standard test sequence, in the assigned format (remote, online or traditional, in-person).

Examiners' scoring decisions were used in the present analysis for any test that required immediate scoring (to determine reaching basals and ceilings/discontinues, for example), in order to determine whether examiners' decisions were affected by format. That is, any test that needed immediate scoring utilized immediate scoring, and each was checked afterward for accuracy by the Project Director, in order to determine if the administration format affected how accurate these scoring decisions were. All tests that did not require immediate scoring were scored after the fact by the Project Director, and all raw scores were entered into the WJ IV online scoring program, regardless of administration format, in the same manner. Data were reviewed for quality by examining any potential outliers and data entry mistakes (e.g., scores that fell outside of the possible range), and entry errors were corrected.

Group equivalence was first explored by comparing the CogAT scores (all of which were collected by traditional administration of the CogAT) of the two different groups (online versus traditional administration of the WJ IV groups). Then, finally, all cluster and test standard/scaled scores were compared between groups to determine whether there is an effect of format on scores.

## Results

Table 2 reports the means and standard deviations of scores on the CogAT nonverbal score and the WJ IV cluster and test scores for each format and for the sample as a whole. Given the close similarity of the demographic characteristics and the balancing by age and gender for the two groups, with random assignment, there would be no expectation of large or systematic differences in scores between the groups.

Table 2
Descriptive Statistics for the CogAT and WJ IV Cluster and Test Scores by Administration Format

| | Traditional, In-Person Administration | | Remote, Online Administration | | Total Sample | |
|---|---|---|---|---|---|---|
| **Test/Cluster** | **Mean** | ***SD*** | **Mean** | ***SD*** | **Mean** | ***SD*** |
| **CogAT** | 103.14 | 16.532 | 103.96 | 15.059 | 103.55 | 15.785 |
| **WJ IV Cognitive** | | | | | | |
| **General Intellectual Ability** | 98.07 | 16.551 | 97.12 | 14.961 | 97.59 | 15.750 |
| ***Gf-Gc* Composite** | 98.30 | 17.991 | 96.82 | 14.673 | 97.56 | 16.399 |
| **Comp-Knowledge** | 94.31 | 16.037 | 94.90 | 12.062 | 94.60 | 14.162 |
| **Fluid Reasoning** | 101.69 | 18.226 | 98.91 | 16.073 | 100.30 | 17.204 |
| **Short-Term Working Memory** | 100.13 | 14.493 | 99.10 | 16.121 | 99.61 | 15.305 |
| **Cognitive Efficiency** | 100.75 | 14.299 | 97.99 | 14.463 | 99.37 | 14.418 |
| **Oral Vocabulary** | 97.14 | 16.524 | 97.22 | 12.106 | 97.18 | 14.454 |
| **Number Series** | 102.19 | 17.237 | 100.69 | 15.115 | 101.44 | 16.194 |
| **Verbal Attention** | 100.43 | 14.721 | 102.22 | 14.993 | 101.32 | 14.854 |
| **Letter-Pattern Matching** | 101.10 | 15.297 | 99.44 | 14.496 | 100.27 | 14.894 |
| **Phonological Processing** | 93.85 | 15.783 | 93.20 | 14.972 | 93.52 | 15.354 |
| **Story Recall** | 95.44 | 13.849 | 92.62 | 15.218 | 94.03 | 14.588 |
| **Visualization** | 98.72 | 15.353 | 97.39 | 13.193 | 98.06 | 14.304 |
| **General Information** | 93.58 | 15.233 | 94.41 | 11.992 | 93.99 | 13.686 |
| **Concept Formation** | 100.41 | 18.304 | 97.39 | 16.488 | 98.90 | 17.453 |
| **Numbers Reversed** | 99.54 | 14.503 | 96.57 | 16.317 | 98.05 | 15.476 |
| **WJ IV Achievement** | | | | | | |
| **Broad Reading** | 98.73 | 14.418 | 99.38 | 14.203 | 99.06 | 14.284 |
| **Broad Mathematics** | 98.17 | 15.338 | 100.08 | 13.096 | 99.13 | 14.259 |
| **Broad Writing** | 105.18 | 13.930 | 107.45 | 14.641 | 106.31 | 14.306 |
| **Letter-Word Identification** | 99.79 | 13.615 | 100.86 | 13.612 | 100.33 | 13.596 |
| **Applied Problems** | 101.45 | 18.786 | 103.08 | 15.027 | 102.26 | 16.995 |
| **Spelling** | 100.20 | 12.835 | 102.37 | 13.537 | 101.28 | 13.206 |
| **Passage Comprehension** | 95.14 | 13.939 | 95.13 | 12.975 | 95.13 | 13.438 |
| **Calculation** | 97.08 | 12.473 | 98.22 | 12.902 | 97.65 | 12.677 |
| **Writing Samples** | 111.22 | 16.715 | 112.38 | 16.692 | 111.80 | 16.679 |
| **Word Attack** | 102.18 | 17.575 | 101.40 | 15.616 | 101.79 | 16.594 |
| **Oral Reading** | 97.98 | 15.342 | 99.38 | 13.186 | 98.68 | 14.292 |
| **Sentence Reading Fluency** | 98.58 | 15.670 | 99.60 | 16.558 | 99.09 | 16.094 |
| **Math Facts Fluency** | 97.49 | 14.769 | 99.00 | 14.809 | 98.25 | 14.777 |
| **Sentence Writing Fluency** | 100.82 | 14.240 | 102.98 | 16.666 | 101.90 | 15.510 |
| ***N*** | 120 | | 120 | | 240 | |

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

Table 3 shows, for the CogAT and each WJ IV cluster and test, the *t* value associated with the format as a predictor, its related *p* value, and its effect size (Cohen's *d*).

Table 3
Significance and Effect Size of Remote, Online Format on the CogAT and the Cluster and Test Scores of the WJ IV

| Test/Cluster | *T* | *p* | Effect Size |
|---|---|---|---|
| **CogAT** | -.400 | 0.689 | -0.052 |
| **General Intellectual Ability** | .466 | 0.641 | 0.060 |
| ***Gf-Gc* Composite** | .700 | 0.485 | 0.090 |
| **Comp-Knowledge** | -.323 | 0.747 | -0.042 |
| **Fluid Reasoning** | 1.255 | 0.211 | 0.162 |
| **Short-Term Working Memory** | .516 | 0.606 | 0.067 |
| **Cognitive Efficiency** | 1.486 | 0.139 | 0.192 |
| **Oral Vocabulary** | -.040 | 0.968 | -0.006 |
| **Number Series** | .717 | 0.474 | 0.093 |
| **Verbal Attention** | -.934 | 0.351 | -0.120 |
| **Letter-Pattern Matching** | .862 | 0.39 | 0.111 |
| **Phonological Processing** | .327 | 0.744 | 0.042 |
| **Story Recall** | 1.504 | 0.122 | 0.194 |
| **Visualization** | .714 | 0.476 | 0.093 |
| **General Information** | -.471 | 0.638 | -0.061 |
| **Concept Formation** | 1.341 | 0.181 | 0.173 |
| **Numbers Reversed** | 1.493 | 0.137 | 0.192 |
| **Broad Reading** | .352 | 0.725 | -0.045 |
| **Broad Mathematics** | -1.034 | 0.302 | -0.134 |
| **Broad Writing** | -1.233 | 0.219 | -0.159 |
| **Letter-Word Identification** | -.607 | 0.544 | -0.079 |
| **Applied Problems** | -.740 | 0.46 | -0.096 |
| **Spelling** | -1.272 | 0.204 | -0.165 |
| **Passage Comprehension** | .010 | 0.992 | 0.001 |
| **Calculation** | -.695 | 0.488 | -0.090 |
| **Writing Samples** | -.537 | 0.592 | -0.069 |
| **Word Attack** | .365 | 0.715 | 0.047 |
| **Oral Reading** | -.754 | 0.452 | -0.098 |
| **Sentence Reading Fluency** | -.489 | 0.626 | -0.063 |
| **Math Facts Fluency** | -.787 | 0.432 | -0.102 |
| **Sentence Writing Fluency** | -1.073 | 0.285 | -0.139 |

*Note.* A positive effect size indicates higher scores with traditional, in-person administration.

No cluster or test had either a significant (at the $p < .05$ level) difference between administrations or an effect size of administration format that exceeded the pre-established criterion of 0.20. Therefore, there does not seem to be a statistically significant effect of the online, remote administration format on examinees' scores.

## Discussion

The present study aimed to evaluate the equivalence of an online, remote administration procedure to traditional, in-person administration of the *Woodcock-Johnson IV* cognitive ability and academic achievement tests. It should be noted that the present study utilized a general school sample, rather than a sample from a specific clinical population, so generalizability is limited. There were no exhibited method effects for the different modes of administration in this nonclinical sample. Different clinical samples, however, may have specific interaction effects with the digital platform for the online, remote administration, which should be considered when interpreting scores in clinical populations.

The present study showed negligible effect sizes (below the 0.20 threshold) and no significant differences (at the $p < 0.05$ level) between administration modes. While more data should be collected to compare the two administration methods, especially with specific clinical populations, the present study suggests that the scores elicited by the two different administration methods are equivalent and interchangeable, and as such all the WJ IV normative and psychometric (reliability, validity, utility) research can be applied confidently to the new online, remote administration of the tests. It should be noted that the online, remote administration of the WJ IV in the present study adhered to very specific, rigorous protocols (with regard to digital specifications, specific training for examiners and proctors, and standardized administration modifications), manualized by PresenceLearning, Inc. The equivalence of remote, online administration that does *not* adhere to these protocols has not been determined.

## References

Lohman, D. F., & Hagen, E. P. (2001). *Cognitive Abilities Test,* Form 6. Itasca, IL: Riverside Publishing.

Schrank, F. A., McGrew, K. S., & Mather, N. M. (2014). *Woodcock-Johnson IV.* Itasca, IL: Riverside Publishing.