# Item Fairness Evaluation of the *Woodcock-Johnson® V* Norming Items

Hyeonjoo Oh, Huan Liu, JP Kim
Riverside Insights®

Unpublished manuscript.
Expanded Version of the Poster Presented at the Annual Meeting of the National Association of School Psychologists, New Orleans, Louisiana. February 14–17, 2024.

## Abstract

In standardized assessments, detecting potentially biased items is important to ensure that tests are fair and unbiased for all examinees. To learn about the concept of item fairness and its evaluation process, differential item functioning (DIF) is introduced. This paper provides the DIF results for the *Woodcock-Johnson® V* (WJ V™) tests. Items with stable and reliable item statistics and without DIF were selected for inclusion in the WJ V publication form after a careful review by content experts and test authors.



Published in Itasca, Illinois.

## Introduction

Diversity, equity, and inclusion (DEI) is a critical area of focus in many fields of educational and psychological measurement that use standardized tests to assess an individual's cognitive abilities, personality, traits, and other aspects. To address DEI concerns related to educational and psychological measurement, it is important to select tests that have been validated across diverse populations, regardless of gender, race, ethnicity, and cultural and linguistic backgrounds, and to carefully review test items to ensure that they are inclusive and do not contain any discriminatory language or assumptions.

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p. 51), "Characteristics of the test itself that are not related to the construct being measured . . . may sometimes result in different meanings for scores earned by members of different identifiable subgroups." One such characteristic is the bias in item difficulty among subgroups, referred to as differential item functioning (DIF). DIF is observed when different groups of people with the *same* underlying ability level respond *differently* to a particular test item. For example, if both male and female students with similar math abilities were asked to solve a math problem using fractions, but the male students perform differently on this item than the female students did, then this item may exhibit DIF based on gender.

It's important to identify biased items with DIF to ensure that tests are fair and unbiased for all examinees. If significant differences are found, it suggests that the item may be biased or unfair for one group compared to another, potentially indicating issues with the validity or fairness of the test. Researchers and test developers use DIF analysis results to identify and address potential sources of bias in assessments, thereby enhancing the fairness and validity of the testing process.

The objectives of this paper are: (1) to introduce the concept of DIF to the school psychology community and (2) to evaluate item fairness through DIF in the *Woodcock-Johnson V* (WJ V) tests. Items with stable and reliable item statistics and without DIF were selected for inclusion in the WJ V publication form after a careful review of the item analyses (e.g., item difficulty review, item discrimination review, item fit statistics review based on the item response theory) and DIF analyses.

## The *Woodcock-Johnson V* Tests

The WJ V, which is currently in development for digital administration, represents a comprehensive revision of the *Woodcock-Johnson IV* (WJ IV®; McGrew, Laforte, & Schrank, 2014), incorporating theoretical, structural, interpretive, and digital enhancements. Aligned with the contemporary Cattell-Horn-Carroll (CHC) theory of cognitive abilities (Schneider & McGrew, 2018), the WJ V is designed to assess general intelligence; broad and narrow cognitive abilities; reading, mathematics, and writing achievement; oral language proficiency; and other cognitive and linguistic skills pertinent to academic success. Drawing from insights in cognitive, neurocognitive, and developmental psychology, as well as research in reading, writing, and mathematics, and user feedback, this revision offers administration and interpretive options tailored to modern assessment demands. Moreover, the transition of the WJ V to a digital testing platform signifies a significant advancement in its accessibility and usability (Laforte, Dailey, & McGrew in press).

The CHC model is a comprehensive psychological theory outlining the structure of human cognitive abilities. According to CHC theory, intelligence is multidimensional and functionally integrated. The theory posits that the dimensions of intelligence can be measured, studied, and understood in terms of their shared and separate antecedents, correlations, and causal effects (Schneider & McGrew, 2018). The WJ V has two co-normed batteries—the WJ V Tests of Cognitive Abilities (WJ V COG) and the WJ V Tests of Achievement (WJ V ACH)—as well as the WJ V Virtual Test Library (WJ V VTL), which is a collection of tests that can be used on its own or in conjunction with the WJ V COG and/or the WJ V ACH. This structure offers examiners the flexibility to use the batteries and VTL independently or in any combination to meet diverse assessment needs (Laforte, Dailey, & McGrew, in press).

## Statistical Methods to Evaluate Item Bias

To conduct DIF analysis, individuals must be familiar with the terminology pertaining to this concept. In a DIF analysis, the group of interest is referred to as the *focal group*, while the group to be compared against is referred to as the *reference group*. In the fraction example mentioned on page 3, the female group is the focal group, and the male group is the reference group. DIF analysis requires matching examinees' abilities between the focal and reference groups. This comparison of matched groups with similar abilities is critical in DIF analysis because it allows the differences in item functioning and the differences between groups to be distinguished. The total test score (or an examinee's estimated latent ability, $\theta$) is used to match the examinees' abilities and is referred to as a *matching criterion* (Dorans & Holland, 1993). For the current study, we used the R package difR (Magis et al., 2010) to perform Mantel-Haenszel (MH) DIF and Standardized P Difference analyses.

### Mantel-Haenszel DIF (MH DIF)

Among DIF methods based on observed scores, the Mantel-Haenszel DIF (MH DIF; Mantel & Haenszel, 1959) was introduced to psychological and educational measurement by Holland and Thayer (1985) to examine the functioning of dichotomously scored items across different groups. The MH DIF analysis involves a 2 (groups) $\times$ 2 (item scores) $\times$ *M* (total score levels) contingency table (see Table 1). The contingency table gives the counts of Right (1) and Wrong (0) responses, and these counts are broken down by the focal and reference groups and the matching criterion.

Table 1. The 2 (Groups) $\times$ 2 (Item Scores) $\times$ *M* (Total Score Levels) Contingency Table

| Group | Item Score | | |
|---|---|---|---|
| | Right | Wrong | Total |
| Focal Group (*f*) | $R_{fm}$ | $W_{fm}$ | $N_{fm}$ |
| Reference Group (*r*) | $R_{rm}$ | $W_{rm}$ | $N_{rm}$ |
| Total Group (*t*) | $R_{tm}$ | $W_{tm}$ | $N_{tm}$ |

*Note*. $R_{rm}$ = Number in reference group at ability level *m* answering the item right; $W_{rm}$ = Number in reference group at ability level *m* answering the item wrong; $R_{fm}$ = Number in focal group at ability level *m* answering the item right; $W_{fm}$ = Number in focal group at ability level m answering the item wrong; and $N_{tm}$ = Number in total group at ability level *m*.

Mantel and Haenszel (1959) provided an estimate of the constant odds ratio ($\alpha_{MH}$), which is an estimate of DIF effect size. The odds of getting the item correct at a given level of the matching criterion are the same in both the focal and reference groups across all *M* levels of the matching criterion.

$$\alpha_{MH} = [\sum_m (R_{rm} W_{fm} / N_{tm})] / [\sum_m (R_{fm} W_{rm} / N_{tm})], m = 1, \ldots, M.$$

Holland and Thayer (1985) converted $\alpha_{MH}$ into a difference in delta (Δ) metric, which is used to scale item difficulty estimates with a mean of 13 and a standard deviation of 4.

$$\text{MH D-DIF} = -2.35 \ln[\alpha_{MH}]$$

Based on these MH D-DIF statistics, items are classified into one of three categories and assigned values of A, B, or C. A negative DIF value implies that, depending on the matching criterion, the focal group has a lower mean item score than the reference group. In contrast, a positive DIF value implies that, depending on the total test score, the reference group has a lower mean item score than the focal group. Table 2 presents the MH D-DIF flagging criteria, a detailed explanation of the categories, and actions for test assembly and score reporting.

Table 2. MH D-DIF Categories and Actions for Test Assembly and Score Reporting

| MH D-DIF Category | Definition | Test Assembly | Score Reporting |
|---|---|---|---|
| A (negligible) | The absolute value of the MH D-DIF is not significantly different from 0 or is less than 1. | Select freely | No action required |
| B (slight to moderate) | The absolute value of the MH D-DIF is significantly different from 1 but is less than 1.5. | If there is a choice, select the item with the smallest MH D-DIF. | No action required |
| C (moderate to large) | The absolute value of the MH D-DIF is significantly different from 1 and is at least 1.5. | Select only if it is essential to meet specifications. Documentation and review panel required | Documentation and review panel required |

*Note*. Positive values (e.g., C+) indicate that the item favors the focal group, whereas negative values (e.g., C–) indicate that the item disadvantages the focal group.

## Standardized P Difference (STD P-DIF)

The STD P-DIF refers to a statistical method used to detect if different groups of examinees respond differently to particular items on a test, even after controlling their overall ability levels. In STD P-DIF analysis, *P* refers to the probability of success on an item (e.g., correctly answering a multiple-choice question). The term *standardized* implies that the analysis adjusts for group differences in overall ability levels. The STD P-DIF is defined as:

$$\text{STD P-DIF} = \frac{\Sigma_m N_{fm} Pfm}{\Sigma_m N_{fm}} - \frac{\Sigma_m N_{fm} P_{rm}}{\Sigma_m N_{fm}},$$

where $P_{fm}$ and $P_{rm}$ are the proportions of examinees who answer correctly over the total number of examinees in the focal and reference groups at score level $m$, $P_{fm} = R_{fm}/N_{fm}$ and $P_{rm} = R_{rm}/N_{rm}$ (Dorans & Holland, 1993). The STD P-DIF is an index that can range from –1 to +1. The STD P-DIF is applied for both dichotomous and polytomous items. Table 3 presents the STD P-DIF flagging criteria, a detailed explanation of the categories, and actions for test assembly and score reporting.

Table 3. STD P-DIF Categories and Actions for Test Assembly and Score Reporting

| STD P-DIF Category | Definition | Test Assembly | Score Reporting |
|---|---|---|---|
| A (negligible) | STD P-DIF values between −.05 and +.05 are considered negligible. | Select freely | No action required |
| B (slight to moderate) | STD P-DIF values between −.10 and –.05 and between .05 and .10 are considered slight to moderate DIF. | If there is a choice, select the item with the smallest STD P-DIF. | No action required |
| C (moderate to large) | STD P-DIF values outside the [−.10, +.10] range are more unusual and should be examined very carefully. | Select only if it is essential to meet specifications. Documentation and review panel required | Documentation and review panel required |

*Note*. Positive values (e.g., C+) indicate that the item favors the focal group, whereas negative values (e.g., C–) indicate that the item disadvantages the focal group.

## Data

The data used in the present study were sourced from the WJ V norming samples collected between late 2021 and 2023. These norming samples were chosen to be, within practical constraints, representative of the U.S. population ranging from 3 years to 85 years and older. For a stable and reliable DIF analysis, the recommended minimum sample size is 100 for the focal group and 400 for the reference group. Table 4 presents the minimum and maximum sample sizes, test mean scores on the W-score scale, and standard deviations for 59 tests across various subgroups in the DIF analysis. With the exception of Letter Writing Fluency (LWRTFL), which evaluated the 4- to 9-year-old age group, all subgroups met the recommended minimum sample size criteria for each test. Due to the narrow age range of LWRTFL, there were insufficient samples available for conducting the DIF analysis.

Table 4. Summary of Sample Size, W-Score Mean, and Standard Deviation for the DIF Analysis Across 59 WJ V Tests

| | | Gender | | Race/Ethnicity | | |
|---|---|---|---|---|---|---|
| | | Male | Female | White | Black | Hispanic |
| *N* | Min | 189 | 195 | 203 | 206 | 278 |
| | Max | 848 | 1,007 | 1,039 | 313 | 402 |
| *M* | Min | 463 | 467 | 465 | 486 | 486 |
| | Max | 523 | 537 | 538 | 521 | 520 |
| *SD* | Min | 6.6 | 5.8 | 6.0 | 6.1 | 6.6 |
| | Max | 64.4 | 64.1 | 62.2 | 60.3 | 68.1 |

## Results

This paper introduces the concept of DIF and outlines detailed procedures for detecting DIF using two commonly employed methods. Table 5 summarizes tests containing C DIF items among the norming items in the WJ V, revealing that only a few dichotomous items were identified as having moderate to large C DIF. None of the polytomous items were flagged as C DIF items. Only 11 out of 2,559 dichotomous items (i.e., 0.43% or 0.39% if the 243 polytomous items are included) were flagged. This serves as excellent evidence of the high quality of the WJ V items created by the authors and selected for the publication form.

The findings across item format align with Bolt's (2000) discovery that multiple-choice (i.e., dichotomous) items exhibit more DIF characteristics than constructed-response (polytomous) items between males and females on SAT math pretest items when the researcher used SIBTEST to detect DIF.

All flagged items underwent a thorough evaluation by the test authors to pinpoint potential sources of bias. As illustrated in Table 5, 6 out of 59 tests flagged only one or two items as C DIF items. In most cases, both MH D-DIF and STD P-DIF identified the same items, although MH D-DIF tended to flag slightly more items as C DIF items compared to STD P-DIF. Appendices A and B provide C DIF information for all 44 tests with dichotomous items and 15 tests with polytomous items, respectively, excluding LWRTFL.

Table 5. Summary of Items Flagged for C DIF for Six WJ V Norming Tests

| Test | Max Points | Mantel-Haenszel D-DIF | | | | | | Standardized P-DIF | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Male vs. Female | | White vs. Black | | White vs. Hispanic | | Male vs. Female | | White vs. Black | | White vs. Hispanic | |
| | | C+ | C– | C+ | C– | C+ | C– | C+ | C– | C+ | C– | C+ | C– |
| Matrices | 50 | 0 | 0 | 1 (2%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Paragraph Reading Comprehension | 74 | 0 | 0 | 0 | 1 (1.4%) | 0 | 0 | 0 | 0 | 0 | 1 (1.4%) | 0 | 0 |
| Picture Vocabulary | 42 | 0 | 0 | 0 | 0 | 0 | 1 (1.7%) | 0 | 0 | 0 | 0 | 0 | 0 |
| Story Comprehension | 36 | 0 | 0 | 0 | 0 | 0 | 1 (1.5%) | 0 | 0 | 0 | 0 | 0 | 1 (1.5%) |
| Sentence Writing Accuracy | 38 | 0 | 0 | 0 | 0 | 1 (0.5%) | 0 | 0 | 0 | 0 | 0 | 1 (0.5%) | 0 |
| Sentence Writing Fluency | 67 | 0 | 0 | 2 (5%) | 1 (2.5%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Note*. Positive values (e.g., C+) indicate that the item favors the focal group, whereas negative values (e.g., C–) indicate that the item disadvantages the focal group.

## Discussion

This paper introduces a concept of DIF and presents detailed procedures for DIF detections by two widely used methods. Differential item characteristics highlight variations in item performance by examinees with the *same* ability across different subgroups. When an item displays DIF, it can lead to unequal total test scores among examinees with similar abilities but from diverse subpopulations, resulting in unfair disadvantages for certain groups. Identifying items with significant DIF and removing them from the test before operational use is vital. After a thorough review of flagged items showing DIF by content experts and WJ V test authors to uncover potential sources of bias, these items are reviewed for possible exclusion from the WJ V publication form.

However, in rare cases, statistically flagged DIF items may remain in the final test forms if, after a thorough sensitivity review, there is no apparent evidence of DIF from a content perspective. This can occur if the flag is due to one or more highly unexpected (i.e., misfitting) examinee responses or a very small sample size for certain subgroups. Identification of DIF is not based solely on statistics; it involves a holistic evaluation that considers both content perspectives and DIF statistics.

This paper aims to provide school psychologists and users of the WJ V tests with an understanding of item fairness, the significance of DIF analysis, and the item evaluation process to foster unbiased and equitable testing practices from the perspective of DEI. When school psychologists and practitioners select a standardized test from the array of available psychological assessments, it is essential to prioritize choosing a test composed of unbiased items that have been pretested and empirically validated. Such tests should be well-designed, thoroughly researched, and normed using a United States representative sample of adequate size.

Using DIF analysis provides researchers with deeper insights into bias issues, facilitating ongoing enhancements in the inclusion and testing of diverse subgroups, including examinees with special needs or examinees with different cultural and linguistic backgrounds, within educational practices. Hence, it is strongly recommended that researchers and practitioners incorporate DIF analysis into the construction of tests or the utilization of test scores for placement, diagnosis, and intervention process.

The current study could not conduct DIF analysis for Asian and other subpopulation groups due to the small sample size. Similarly, DIF results for LWRTFL are not included in this paper because of the limited samples. However, future analyses could explore small-sample DIF for both Asian and other subpopulations as well as for LWRTFL. Additionally, few studies have investigated DIF analysis for special needs groups, such as those with learning disabilities, compared to typical examinees using the WJ family of assessments. Further research is warranted to examine DIF in these special groups alongside typical examinees.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. AERA.

Bolt, D. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement*, 37(4), 307–327.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum.

Holland, P. W., & Thayer, S. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Report RR-85-43). Educational Testing Service.

LaForte, E. M., Dailey, D., & McGrew, K. S. (in press). Technical Manual. Woodcock-Johnson V. Riverside Assessments, LLC.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, 42(3), 847–862.

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. Journal of the National Cancer Institute, 22, 719-748.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical Manual. *Woodcock-Johnson IV*. Riverside Assessments, LLC.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). Guilford Press.

Appendix A. Summary of Items Flagged for C DIF for 44 WJ V Norming Tests With Dichotomous Items (*Continued*)

## Appendices

### Appendix A. Summary of Items Flagged for C DIF for 44 WJ V Norming Tests With Dichotomous Items

| Test | Number of items | Mantel-Haenszel D-DIF | | | | | | Standardized P-DIF | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Male vs. Female | | White vs. Black | | White vs. Hispanic | | Male vs. Female | | White vs. Black | | White vs. Hispanic | |
| | | C+ | C– | C+ | C– | C+ | C– | C+ | C– | C+ | C– | C+ | C– |
| Academic Facts | 55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Academic Vocabulary | 63 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Animal-Number Sequencing | 31 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Analysis-Synthesis | 35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Applied Problems | 55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Block Rotation | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Calculation | 56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Concept Formation | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| General Information | 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Letter-Pattern Matching | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Letter-Word Identification | 90 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Magnitude Comparison | 120 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Matrices | 50 | 0 | 0 | 1 (2%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Memory for Words | 26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Math Problem Identification | 55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Number-Pattern Matching | 90 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Numbers Reversed | 34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Number Sense | 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Number Series | 42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nonsense Word Repetition | 46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| Test | Number of items | Mantel-Haenszel D-DIF | | | | | | Standardized P-DIF | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Male vs. Female | | White vs. Black | | White vs. Hispanic | | Male vs. Female | | White vs. Black | | White vs. Hispanic | |
| | | C+ | C– | C+ | C– | C+ | C– | C+ | C– | C+ | C– | C+ | C– |
| Oral Comprehension | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Oral Vocabulary | 63 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Paragraph Reading Comprehension | 74 | 0 | 0 | 0 | 1 (1.4%) | 0 | 0 | 0 | 0 | 0 | 1 (1.4%) | 0 | 0 |
| Picture Vocabulary | 42 | 0 | 0 | 0 | 0 | 0 | 1 (1.7%) | 0 | 0 | 0 | 0 | 0 | 0 |
| Passage Comprehension | 46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Segmentation | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sentence Repetition | 63 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sound Blending | 74 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sound Deletion | 59 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sound Reversal | 48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sound Substitution | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Spatial Relations | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Spelling | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Spelling of Sounds | 42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sentence Reading Fluency | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Story Comprehension | 36 | 0 | 0 | 0 | 0 | 0 | 1 (1.5%) | 0 | 0 | 0 | 0 | 0 | 1 (1.5%) |
| Sentence Writing Accuracy | 38 | 0 | 0 | 0 | 0 | 1 (0.5%) | 0 | 0 | 0 | 0 | 0 | 1 (0.5%) | 0 |
| Sentence Writing Fluency | 67 | 0 | 0 | 2 (5%) | 1 (2.5%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Understanding Directions | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Visual-Auditory Learning | 120 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Verbal Analogies | 41 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Verbal Attention | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Word Attack | 63 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Word Reading Fluency | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Notes.*
Positive C DIF values suggest that the item favors the focal group (i.e., female, Black, and Hispanic groups), whereas negative C DIF values indicate that the item disadvantages the focal group.

MH D-DIF is applicable only to dichotomous items.

## Appendix B. Summary of Items Flagged for C DIF for 15 WJ V Norming Tests With Polytomous Items

| Test | Number of items | Standardized P-DIF | | | | | |
|---|---|---|---|---|---|---|---|
| | | Male vs. Female | | White vs. Black | | White vs. Hispanic | |
| | | C+ | C– | C+ | C– | C+ | C– |
| Oral Language Samples | 35 | 0 | 0 | 0 | 0 | 0 | 0 |
| Math Facts Fluency | 16 | 0 | 0 | 0 | 0 | 0 | 0 |
| Oral Reading | 28 | 0 | 0 | 0 | 0 | 0 | 0 |
| Phonemic Word Retrieval | 16 | 0 | 0 | 0 | 0 | 0 | 0 |
| Reading Recall | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rapid Letter Naming | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rapid Number Naming | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rapid Phoneme Naming | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rapid Picture Naming | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rapid Quantity Naming | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| Semantic Word Retrieval | 16 | 0 | 0 | 0 | 0 | 0 | 0 |
| Story Recall | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| Symbol Inhibition | 18 | 0 | 0 | 0 | 0 | 0 | 0 |
| Visual Working Memory | 19 | 0 | 0 | 0 | 0 | 0 | 0 |
| Written Language Samples | 35 | 0 | 0 | 0 | 0 | 0 | 0 |