Journal of Psychoeducational Assessment
1–14
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0734282917698303
journals.sagepub.com/home/jpa
SAGE

# Testing for Construct Bias in the Differential Ability Scales, Second Edition: A Comparison Among African American, Asian, Hispanic, and Caucasian Children

**Katherine M. Trundt[1], Timothy Z. Keith[2], Jacqueline M. Caemmerer[2], and Leann V. Smith[2]**

## Abstract

Individually administered intelligence measures are commonly used in diagnostic work, but there is a continuing need for research investigating possible test bias among these measures. One current intelligence measure, the Differential Ability Scales, Second Edition (DAS-II), is a test with growing popularity. The issue of test bias, however, has not been thoroughly investigated with the DAS-II. The current study investigated whether the DAS-II demonstrates systematic construct bias when used with children from three racial and ethnic groups—African American, Asian, and Hispanic—when compared to non-Hispanic Caucasian children. Multigroup confirmatory factor analyses with data from the DAS-II standardization sample were used to assess whether the constructs and measurement of constructs were invariant across groups. Results indicate cross-group internal structure validity in the DAS-II, and thus a lack of construct bias. Minor differences were found, but these differences do not affect the calculation of composite scores on the DAS-II and thus would not result in unfair scoring for the groups involved. Results of this study support the appropriateness of the DAS-II for clinical use with these racial and ethnic groups.

## Keywords

cross-cultural comparison, culture/crosscultural, factor analysis, measurement, cognitive abilities, personality/individual differences, other, intelligence tests, intelligence/cognition

Debate has shadowed the field of intelligence testing since its inception. Given the complicated questions regarding what constitutes intelligence, how it can best be measured, and whether the resulting scores are meaningful and equivalent across different groups, the history of controversy comes as no surprise. Intelligence testing traces back to the late 1800s, and the Binet–Simon Scale is typically considered the first modern intelligence test (Cohen & Swerdlik, 2002).

[1]Del Valle Independent School District, TX, USA
[2]The University of Texas at Austin, TX, USA

**Corresponding Author:**
Timothy Z. Keith, Department of Educational Psychology, The University of Texas at Austin, 254B SZB, 1912 Speedway D5800, Austin, TX 78712-1289, USA.
Email: tzkeith@austin.utexas.edu

Intelligence tests were soon widely distributed in the United States and used for a variety of purposes. Despite warnings regarding the limitations of Binet's test and intelligence testing in general (Binet, Simon, & Kite, 1916), the tests were often administered with little thought about the appropriateness of their use. In particular, early intelligence measures were developed at a time when culture bias and fairness were not yet a part of the zeitgeist; the lack of cultural considerations during the development of the first intelligence tests may account for many of the criticisms pertaining to cultural bias that still exist today.

Critics of early tests and intelligence testing pointed out the possible flaws of the tests when used with students from racial/ethnic minority groups. These included potential language bias when testing those whose primary language is not English, tests being tied to middle class culture, possible differences in the nature of constructs across groups, the lack of inclusion of minority group members in standardization samples, and the likelihood of differential outcomes for groups based on intelligence test scores. These and other criticisms are detailed elsewhere (e.g., Jensen, 1980; Reynolds & Lowe, 2009; Valencia & Suzuki, 2001, Chapter 5). With advances in technology and methodology, the issue of test bias has become conceptualized primarily as that of differential psychometric validity. From this perspective, test bias may be investigated psychometrically as differential test content validity, internal structure validity, and test-criterion validity (in addition to differential reliability). Evidence concerning modern intelligence tests generally supports a lack of cultural bias against American, English-speaking ethnic minority groups on modern cognitive measures (cf. Reynolds & Lowe, 2009; Valencia & Suzuki, 2001).

Valencia and Suzuki (2001) identified investigations into cultural bias for 14 different intelligence measures, most of which have since been revised once or more. Newer versions of many of these well-known intelligence measures have also been subjected to bias analyses, with generally positive results. Evidence suggests, for example, the Wechsler Intelligence Scale for Children–Fifth Edition (Scheiber, 2016b) and the Kaufman Assessment Battery for Children–Second Edition measure the same underlying constructs, equally well, for African American, Hispanic, and Caucasian children (Scheiber, 2016a) in the standardization samples for these two instruments. Analyses of Raven's Advanced Matrices have shown equivalent factor structures for African and non-African engineering students in South Africa (Rushton, Skuy, & Bons, 2004), and there is evidence to support the equivalence of factor structures for African American and Caucasian children for the previous edition of the Woodcock Johnson Cognitive battery (Edwards & Oakland, 2006). The importance of evaluating cognitive measures for cultural bias cannot be emphasized enough. These intelligence tests are among the most popular measures that psychologists administer and are used for very diverse purposes, many of which have life-changing implications.

## Current Study

The Differential Ability Scales, Second Edition (DAS-II, Elliott, 2007) is a revision of the original Differential Ability Scales (DAS; Elliott, 1990). The DAS-II is a popular measure for assessing children and adolescents for a number of reasons, including its ease of use, appeal to young children, use of a general score based on high *g*-loading measures, and availability of a nonverbal composite (Dumont, Willis, & Elliott, 2009). Therefore, it is important to evaluate whether the DAS-II is an appropriate measure of cognitive abilities for children from diverse backgrounds. There is internal structure validity evidence in the research literature for DAS-II standardization data scores (Canivez & McGill, 2016; Elliott, 2007; Keith, Low, Reynolds, Patel, & Ridley, 2010; Keith, Reynolds, Roberts, Winter, & Austin, 2011), and the DAS-II technical manual presents evidence supporting a lack of bias for items and the prediction of academic achievement (Elliott, 2007). The original DAS showed invariance in the measurement of constructs across three racial/ethnic groups (African American, Hispanic, Caucasian; Keith, Quirk, Schartzer, & Elliott, 1999). There is also preliminary research exploring the possibility of construct bias (i.e., determining whether the internal structure validity evidence is consistent across cultural groups): Trundt's (2013) dissertation showed invariance in factor structures for African American, Asian, and Hispanic children compared to two subsamples of Caucasian children from the DAS-II standardization sample.

The current study was designed to investigate whether the DAS-II demonstrates systematic construct bias toward children of any of three racial/ethnic groups: African American, Asian, and Hispanic, as compared to Caucasian children. In particular, we sought to determine (a) whether construct bias is present in the DAS-II toward any of these groups, (b) if so, where this bias exists, and (c) whether the findings would replicate with additional comparison groups. Multigroup confirmatory factor analysis (MG-CFA) using data from the DAS-II standardization sample assessed whether criteria for increasingly strict levels of invariance were met across groups. These analyses were used to determine whether the DAS-II measures the same constructs across groups, and thus test for construct bias across groups. The analyses reported here are similar to those reported by Trundt (2013), although with additional replication subsamples and the addition of another DAS-II subtest.

## Method

### Instrumentation

The DAS-II (Elliott, 2007) is an individually administered test of cognitive abilities for children and adolescents ages 2:6 to 17:11. The current investigation focused on children ages 5 through 17 years who were administered a common battery of tests from the DAS-II. A description of each subtest included in the current analysis is provided in Table 1. The DAS-II yields an overall composite score, lower-level diagnostic "cluster" scores, and specific ability measures.

*Reliability and validity evidence for the DAS-II.* Evidence provided in the technical manual suggests adequate to strong evidence of reliability in the standardization data (Elliott, 2007). For the overall sample, average corrected test–retest reliability coefficients of subtest, cluster, and composite scores range from .63 (one subtest, Recognition of Pictures) to .91 (one composite, General Conceptual Ability) over a retest interval of 1 to 9 weeks. Evidence reported in the manual also provides internal and external evidence of score validity. As already noted, independent evaluations have shown that the DAS-II appears to measure its intended structure for ages 4:0 to 17:11 (Keith et al., 2010).

### Participants

Participants were selected from the DAS-II standardization sample. The standardization sample was stratified according to age, sex, race/ethnicity, parent education level, and geographic region based on data gathered in 2005 by the U.S. Census Bureau (Current Population Survey). The sample included 2,270 children ages 5:0-17:11 from African American, Asian, Hispanic, and Caucasian racial/ethnic groups. Sample sizes and demographic data for each group are shown in Table 2. Because there were many more Caucasian children in the total sample, Caucasian participants were selected at random from the total sample to form four Caucasian subsamples to be equal (or nearly equal) in size to the largest of the other sample sizes. This strategy allowed four comparisons (one analysis and three replications) for each racial/ethnic group (the data from African American, Asian, and Hispanic subsamples compared to each of the four Caucasian subsamples).

Details regarding countries of origin for racial/ethnic minority children, particularly those of Asian and Hispanic descent, were not available. According to the test manual, all children spoke

**Table 1.** Description of the DAS-II Subtests.

| Subtest | Description |
|---|---|
| Word Definitions | Child defines words presented orally by the examiner. |
| Verbal Similarities | Child explains how three named things or concepts go together. |
| Matrices | Child solves visual puzzles by selecting image missing from a 2 × 2 or 3 × 3 matrix. |
| Sequential and Quantitative Reasoning | Child determines which image completes a sequence of pictures, numbers, or geometric figures. |
| Pattern Construction | Child uses wooden blocks, plastic blocks, or flat tiles to recreate constructions made by the examiner or presented in pictures. |
| Recall of Designs | Child reproduces line drawings from memory after viewing design for 5 s. |
| Recognition of Pictures | Child views images for a few seconds before being asked to select images viewed from a set of pictures. |
| Recall of Objects Immediate | Child is taught names of 20 pictures immediately before asked to recall as many pictures as possible within a time limit. |
| Recall of Objects Delayed | Child is asked to recall pictures from Recall of Objects-Immediate after a 10- to 30-min delay (with pictures not reshown) |
| Recall of Digits Forward | A standard digit recall forward task. |
| Recall of Digits Backward | A standard digits-reversed task. |
| Recall of Sequential Order | Child recalls body parts and other objects presented orally by the examiner, in a different, pre-specified order. |
| Speed of Information Processing | Child marks the circle with the most parts in each row as quickly as possible. |
| Rapid Naming | Child names colors or images as quickly as possible. |

*Note.* DAS-II = Differential Ability Scales, Second Edition.

English, with bilingual children included only if English was reported to be the child's primary language. In addition, all children were able to communicate verbally at a level consistent with their age.

## *Procedure*

The current study used MG-CFA to investigate whether the DAS-II demonstrates construct bias toward children of any of three racial/ethnic groups—African American, Asian, or Hispanic—with Caucasian children as the comparison group. This was accomplished by testing for measurement invariance across groups: evaluating whether the underlying constructs measured by the DAS-II vary based on group membership. Measurement invariance is frequently evaluated using MG-CFA (e.g., see Scheiber, 2016a, 2016b; for the method, see Cheung & Rensvold, 2002; Keith, 2015, Chapter 19).

## *Analysis*

*Descriptive statistics.* Subsamples were selected using IBM SPSS Statistics, Version 23.0 (IBM Corp., 2014). Raw data were analyzed via SPSS Amos, Version 23.0 (Arbuckle, 2014) using MG-CFA. Missing data were minimal, with only eight missing individual subtest standard scores

**Table 2.** Demographic Characteristics of Study Participants.

| Variable | African American | Asian | Hispanic | Caucasian 1 | Caucasian 2 | Caucasian 3 | Caucasian 4 | Total |
|---|---|---|---|---|---|---|---|---|
| *n* | 407 | 98 | 432 | 432 | 432 | 432 | 341 | 2,574 |
| Sex | | | | | | | | |
| Male | 198 | 39 | 218 | 222 | 218 | 222 | 172 | 1,289 |
| | 48.6% | 39.8% | 50.5% | 51.4% | 50.5% | 51.4% | 50.4% | 49.9% |
| Female | 209 | 59 | 214 | 210 | 214 | 210 | 169 | 1,285 |
| | 51.4% | 60.2% | 49.5% | 48.6% | 49.5% | 48.6% | 49.6% | 50.1% |
| Age | | | | | | | | |
| 5:0-7:11 | 91 | 26 | 114 | 101 | 81 | 107 | 74 | 594 |
| | 22.4% | 26.5% | 26.4% | 23.4% | 18.8% | 24.8% | 21.7% | 23.1% |
| 8:0-10:11 | 95 | 20 | 104 | 109 | 99 | 87 | 77 | 591 |
| | 23.4% | 20.4% | 24.1% | 25.2% | 22.9% | 20.1% | 22.6% | 23.0% |
| 11:0-13:11 | 95 | 21 | 98 | 93 | 103 | 93 | 92 | 595 |
| | 23.3% | 21.4% | 22.7% | 29.9% | 23.8% | 21.5% | 27.0% | 23.1% |
| 14:0-17:11 | 126 | 31 | 116 | 129 | 149 | 145 | 98 | 794 |
| | 31.0% | 31.6% | 26.9% | 29.9% | 34.5% | 33.6% | 28.7% | 30.8% |

in the total sample (out of 14 individual subtest standard scores for 2,270 children). For follow-up analyses examining modification indices through Amos, moment matrices (covariances and means) were analyzed rather than raw data (e.g., Graham & Coffman, 2012).

*Data analysis.* We evaluated whether the DAS-II is biased against any of these racial/ethnic groups by testing for construct bias or measurement invariance across racial/ethnic groups. The model used for these analyses is illustrated in Figure 1, with subtests limited to those administered to all children within this specified age range. Group comparisons were conducted in a pairwise fashion, so as to compare Caucasian subsamples individually with African American, Asian, and Hispanic subsamples. Procedures were replicated with three additional Caucasian comparison subsamples to explore the reliability of initial findings.

Three levels of invariance were tested: configural, metric, and intercept. Each level of invariance was progressively stricter than the previous level, meaning that with each step of invariance testing, additional equality constraints were imposed across groups. If invariance was not present across groups, partial invariance was tested (described further in the results). In this formulation, a lack of invariance suggests that the test measures different constructs across groups, and thus the presence of construct bias.

*Configural invariance*. Analysis at this level assessed whether the constructs have the same configuration, or pattern of factor loadings, across groups. To test for configural invariance, the same factor structure was modeled for both groups but the parameters estimated across models (factor loadings, intercepts, variances, etc.) were free to vary. Means of all latent variables were fixed to zero, while the intercepts for the measured variables were freely estimated for all groups. Measures of overall model fit were considered before moving to the next step of the analyses.

*Metric invariance*. This level of invariance examined whether the relation of the measured variables to the latent variables, or the scale of the latent variables, is the same across groups. Unstandardized factor loadings were constrained to be equal across groups. Comparisons of model fit were made; a significant decline in model fit from the configural model to the metric

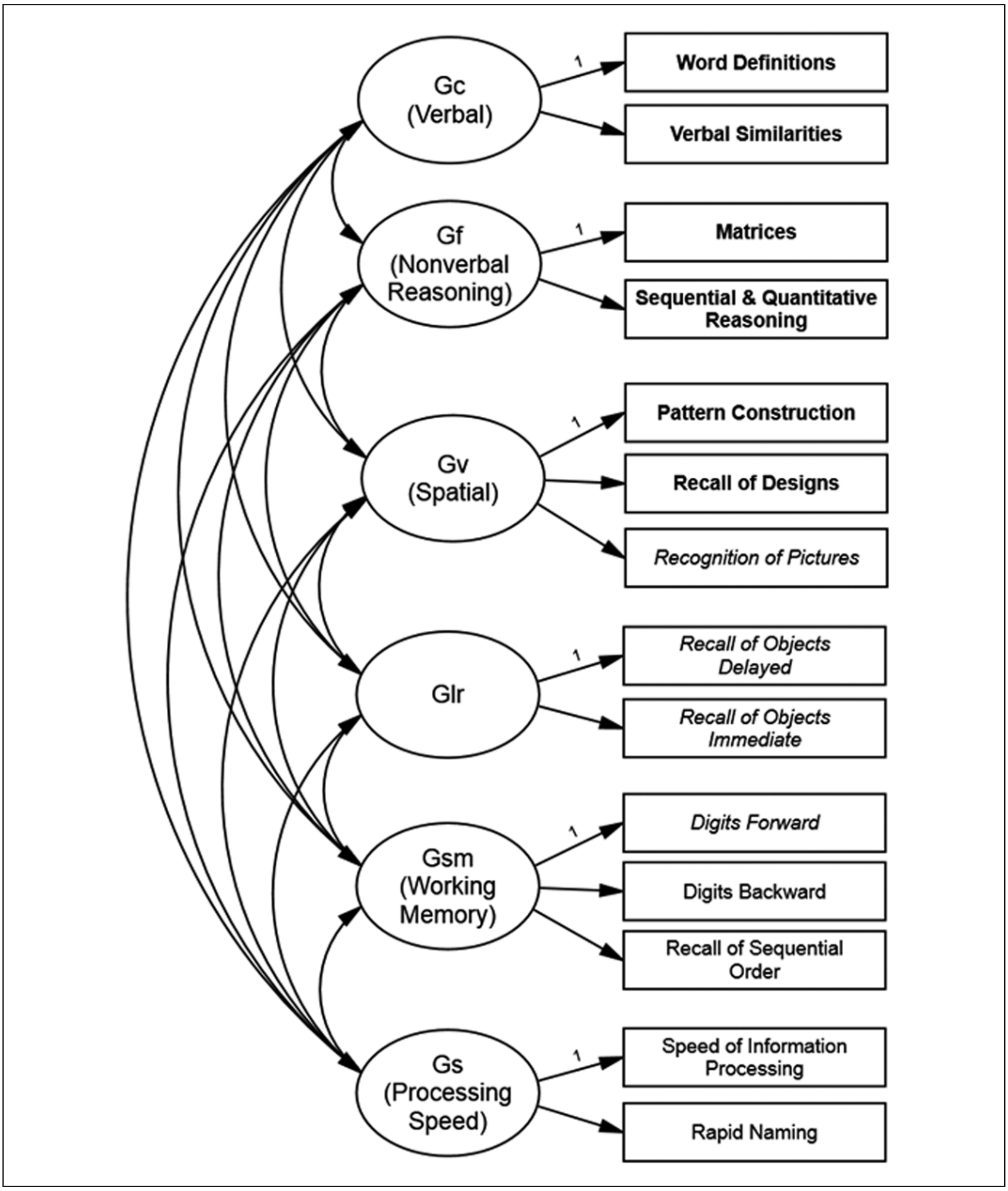

**Figure 1.** First-order factor structure of the DAS-II, ages 5 to 17 years.
*Note*. Roman-text subtests are used to create composite scores designed to measure the constructs (factors) shown; italicized subtests are not used to calculate composite scores. Composite names are shown in parentheses underneath the factor names. There is no composite score corresponding to the Glr factor. Bolded subtests are used in calculation of the General Conceptual Ability score. Error terms for subtests are not shown in order to simplify the figure. DAS-II = Differential Ability Scales, Second Edition.

model suggests a lack of support for metric invariance. As Keith and Reynolds (2012) explained, if factor loadings are the same across groups (metric invariance model supported) then a one-unit increase in a latent variable will result in the same increase in the measured variables across groups.

*Intercept invariance.* Invariance at this level of analysis suggests that, given equal scores on latent factors, the intercepts of the measured variables are the same across groups. Factor means were allowed to vary across groups (these were set to zero in previous steps), and all corresponding subtest intercepts were constrained to be equal across groups (previously these were estimated freely). Differences in intercepts—and therefore means—on measured variables are indicative of a systematic advantage for one group over another, or different starting levels for one group versus the other on the measured variables (subtests) (Keith & Reynolds, 2012). If this level of invariance is supported, then differences in latent (factor) means should account for any differences across groups in the observed variables' (i.e., subtest) scores. If intercept invariance was not achieved, partial intercept invariance was tested, a process which is equivalent to assessing content bias at the subtest level.

*Model fit.* Model fit was assessed at each level of invariance testing, and additional model fit comparisons were made at the metric and intercept invariance levels. As recommended by methodologists, several fit indices were used to evaluate model fit (Hu & Bentler, 1999). For assessing the fit of single models in the current study, the root mean square error of approximation (RMSEA) and the comparative fit index (CFI) were used (e.g., Boomsma, 2000; McDonald & Ho, 2002). RMSEA values below .05 (Browne & Cudeck, 1993) or .06 (Hu & Bentler, 1999) indicate good fit; CFI values over .95 suggest good fit; and CFI values over .90 represent reasonable fit (Hu & Bentler, 1999).

Model fit comparisons were necessary to determine whether the additional constraints imposed at each level of invariance testing significantly degraded the fit of the model to the data. The degree of invariance between two nested models is often assessed using the Likelihood Ratio Test—the difference in chi-square between two models ($\Delta\chi^2$; Cheung & Rensvold, 2002). However, evidence suggests the $\chi^2$ statistic is highly sensitive to sample size and it may not be practically useful to compare competing models when testing for invariance (e.g., Cheung & Rensvold, 2002). Based on simulation research, Cheung and Rensvold (2002) argued that a decrease in CFI (ΔCFI) of greater than −.01 across models suggests a lack of invariance across groups and bias may be present.

## Results

Goodness-of-fit indices for the invariance steps for all models are shown in Table 3. The invariance steps will be described in more detail for the first comparison.

### Data Analysis

#### African American–Caucasian

*Step 1: Configural invariance.* The patterns of loadings were set to be the same for both groups. Means of all latent variables were fixed to zero, while the intercepts for the measured variables were freely estimated for all groups. Results of the initial configural invariance comparison between the African American and Caucasian 1 (the Caucasian subsample used for the first group comparisons across all groups) groups suggested excellent model fit (see Table 3). In particular, CFI was larger than .95, and RMSEA was below .05. For both groups, factor loadings were generally high and factor correlations reasonable.

*Step 2: Metric invariance (weak factorial invariance).* To assess for metric invariance, unstandardized factor loadings were constrained to be equal across groups. These additional constraints imposed to assess metric invariance between the African American and Caucasian 1 samples resulted in a ΔCFI of −.004, which is smaller than the −.01 threshold and suggests model fit did

**Table 3.** Fit Indices and Comparisons for All Models.

| Group | Invariance step | $\chi^2$ | *df* | CFI | ΔCFI[a] | RMSEA[b] |
|---|---|---|---|---|---|---|
| African American vs. Caucasian 1 | Configural | 242.413 | 124 | .971 | — | .048 |
| | Metric | 263.695 | 132 | .967 | −.004 | .049 |
| | Intercept | 313.761 | 140 | .957 | −.0104 | .055 |
| | Partial intercept | 295.210 | 139 | .961 | −.006 | .052 |
| African American vs. Caucasian 2 | Configural | 234.055 | 124 | .979 | — | .047 |
| | Metric | 268.698 | 132 | .974 | −.005 | .049 |
| | Intercept | 327.250 | 140 | .964 | −.0097 | .057 |
| | Partial intercept | 300.213 | 139 | .969 | −.005 | .052 |
| African American vs. Caucasian 3 | Configural | 226.694 | 124 | .978 | — | .044 |
| | Metric | 252.533 | 132 | .974 | −.004 | .047 |
| | Intercept | 302.922 | 140 | .965 | −.0094 | .052 |
| | Partial intercept | 284.603 | 139 | .968 | −.006 | .049 |
| African American vs. Caucasian 4 | Configural | 217.583 | 124 | .976 | — | .045 |
| | Metric | 236.006 | 132 | .973 | −.003 | .046 |
| | Intercept | 284.716 | 140 | .963 | −.0104 | .053 |
| | Partial intercept | 265.902 | 139 | .967 | −.006 | .049 |
| Asian vs. Caucasian 1 | Configural | 175.240 | 124 | .978 | — | .040 |
| | Metric | 193.107 | 132 | .974 | −.004 | .042 |
| | Intercept | 220.043 | 140 | .966 | −.008 | .047 |
| Asian vs. Caucasian 2 | Configural | 166.894 | 124 | .988 | — | .037 |
| | Metric | 178.951 | 132 | .987 | −.001 | .037 |
| | Intercept | 204.624 | 140 | .982 | −.005 | .042 |
| Asian vs. Caucasian 3 | Configural | 159.543 | 124 | .988 | — | .033 |
| | Metric | 181.565 | 132 | .983 | −.005 | .038 |
| | Intercept | 207.467 | 140 | .977 | −.006 | .042 |
| Asian vs. Caucasian 4 | Configural | 150.390 | 124 | .988 | — | .031 |
| | Metric | 167.139 | 132 | .984 | −.004 | .035 |
| | Intercept | 192.087 | 140 | .977 | −.007 | .041 |
| Hispanic vs. Caucasian 1 | Configural | 205.449 | 124 | .979 | — | .040 |
| | Metric | 213.388 | 132 | .979 | .000 | .038 |
| | Intercept | 237.698 | 140 | .975 | −.004 | .040 |
| Hispanic vs. Caucasian 2 | Configural | 197.091 | 124 | .986 | — | .037 |
| | Metric | 206.744 | 132 | .985 | −.001 | .037 |
| | Intercept | 232.370 | 140 | .982 | −.003 | .040 |
| Hispanic vs. Caucasian 3 | Configural | 197.091 | 124 | .986 | — | .037 |
| | Metric | 206.744 | 132 | .985 | −.001 | .037 |
| | Intercept | 232.370 | 140 | .982 | −.003 | .040 |
| Hispanic vs. Caucasian 4 | Configural | 180.631 | 124 | .985 | — | .034 |
| | Metric | 185.780 | 132 | .986 | .001 | .033 |
| | Intercept | 205.261 | 140 | .974 | −.012 | .035 |

*Note.* CFI = comparative fit index; RMSEA = root mean square error of approximation.

[a] When ΔCFI values are close to −.01 four decimal places are shown.

[b] RMSEA corrected for two groups.

not significantly degrade, thus providing support for metric invariance between these two groups. In other words, it appears that the scale—or the relations between the factors and the subtests—is the same for both groups.

*Step 3: Intercept invariance (strong factorial invariance).* Factor means were allowed to vary across groups, and all corresponding subtest intercepts were constrained to be equal across groups. Subtest intercepts may be considered subtest means with the effects of the latent factor controlled. Although overall model fit was still adequate to good based on CFI and RMSEA (see Table 3), the ΔCFI for this step was larger than the threshold value (−.0104), which suggests a lack of intercept invariance across the two groups. Additional analysis showed the primary reason for this misfit was due to the Recall of Digits Forward subtest (Gsm factor). When this intercept was allowed to vary across groups, the ΔCFI value reduced to −.006, thus supporting partial intercept invariance. Partial intercept invariance between the African American and Caucasian 1 samples suggests any differences in latent (factor) means appear to account adequately for all mean differences on subtest scores, *except* for those differences in scores on the Recall of Digits Forward subtest. Comparison of intercepts for the Recall of Digits Forward subtest showed that, within levels of Gsm, African American students scored higher than Caucasian students (by 2.98 points, on average); the subtest scores of African American students are thus higher than expected given their performance on the Gsm/Working Memory factor.

*Replication comparisons: African American–Caucasian.* Table 3 also shows the fit statistics for the three replications, comparing the African American sample with the Caucasian 2, 3, and 4 subsamples ($n$ = 432, 432, and 407, respectively). As shown in the table, the metric invariance step fit well for all three comparisons, which suggests that the DAS-II measures the same constructs for African American and Caucasian children aged 5 to 17 years. For all three replications, the intercept invariance comparison was close to the −.01 threshold for supporting invariance, and was slightly below the threshold for Comparisons 2 and 3 and slightly above for Comparison 4. Because this value was close to −.01 the source of misfit and partial intercept invariance were investigated for all three replication comparisons. In all subsample comparisons, Recall of Digits Forward was the source of misfit, and partial intercept invariance with this intercept free to vary was supported for all replications. Recall of Digits Forward is not used in any composites for the DAS-II, thus the consequences of the lack of intercept invariance for this subtest are minor and do not influence inferences about individuals' levels of general or specific cognitive abilities.

Intercept invariance (and intercept bias) is important for measures like the DAS-II where subtests scores are used to create composites, which are in turn used to make inferences about individuals' levels of general or specific cognitive abilities. If Recall of Digits Forward were used to create the General Conceptual Ability (a measure of overall intelligence) or the Working Memory composite scores, then those scores would be artificially inflated for one group versus the other. The fact that Recall of Digits Forward is not used in any composites for the DAS-II, however, means that the consequences of the lack of intercept invariance for this test are minor.

*Asian Caucasian.* As shown in Table 3, the configural invariance model fit well for the initial comparison between the Asian ($n$ = 98) and Caucasian 1 subsample, and the model fit did not degrade (ΔCFI) for metric invariance or intercept invariance steps. In the configural invariance step, however, the standardized factor loading of the Recall of Objects Immediate subtest on the Glr factor in the Asian sample was larger than 1 (1.11), an impossibility (all other factor loadings, however, were reasonable). This problem is likely related to the relatively small sample for the Asian group. The problem disappeared in the next two invariance steps once factor loadings were constrained to be equal across groups. For this reason, and because supplemental analyses suggested an equally good series of models without this factor, we here report the fit of the models with this factor retained. This pattern was replicated in the subsequent comparisons with the remaining three Caucasian subsamples as well. The findings do suggest, however, that these two tests (Recall of Objects Immediate and Recall of Objects Delayed) may not form a coherent factor for this group. Because there is no Glr composite score reported for the DAS-II, however, this finding has few implications for the practical use of the DAS-II with Asian children.

*Hispanic Caucasian.* As shown in Table 3, the configural invariance model fit well for the initial comparison between the Hispanic ($n$ = 432) sample and first Caucasian subsample. All factor loadings were moderate to large, and the factor correlations were of reasonable magnitude. Model fit was stable (ΔCFI) for the metric invariance and intercept invariance steps. This pattern was replicated in the subsequent comparisons with the remaining three Caucasian subsamples, suggesting that the DAS-II measures the same factors for Hispanic children as it does for Caucasian children.

## Discussion

The DAS-II is an individual intelligence test commonly used to assess the cognitive abilities of children and adolescents. Although there is evidence to support the internal structure validity of the DAS-II (e.g., Keith et al., 2010), and evidence to support differential internal structure validity (and thus a lack of construct bias) across racial/ethnic groups for the earlier version of the test (Keith et al., 1999), evidence concerning construct bias is limited for this latest version of the test (Trundt, 2013). The purpose of the current study was to investigate whether the DAS-II demonstrates systematic construct bias toward African American, Asian, and Hispanic children compared with Caucasian children. The current study was designed to determine (a) whether construct bias is present in the DAS-II toward any of these groups, (b) if so, where this bias exists, and (c) whether the findings would replicate with additional comparison groups. To fulfill this purpose, MG-CFA was used to test for measurement invariance (and construct bias) across groups using data from the DAS-II standardization sample. This methodology incrementally tested whether criteria for increasingly strict levels of invariance were met across groups (e.g., Keith & Reynolds, 2012; Meredith, 1993).

The results of these analyses provide strong support for internal structure validity and a lack of construct bias for Asian and Hispanic children, as compared to Caucasian children. In other words, the underlying attributes and constructs measured by the DAS-II were statistically indistinguishable across these groups, leading to the conclusion that the test measures the same constructs, equally well, for each of these racial/ethnic groups compared to Caucasian children.

Differences in measurement were found for African American, as compared with Caucasian, children, however. In two of the four comparisons, African American children showed differences in intercepts for one test, Recall of Digits Forward, compared with Caucasian children, with this difference favoring African American children. What this means is African American children scored higher on the Recall of Digits Forward subtest than would be expected given their level of Gsm (Working Memory) as compared to Caucasian children. In other words, this subtest may show slightly inflated scores for African American children once Gsm is controlled. As already noted, intercept invariance (and intercept bias) is important for measures like the DAS-II where subtests scores are used to create composites, which are in turn used to make inferences about individuals' levels of general or specific cognitive abilities. If the Recall of Digits Forward test were used in the calculation of composite scores (e.g., General Conceptual Ability or the Working Memory composite score), then these scores would also likely be over-estimates. Recall of Digits Forward is not used in any composite scores, however, and thus the consequences of this difference are minimal. Recall of Digits Backward and Recall of Sequential Order are the two subtests used to create a Working Memory composite score on the DAS-II, and these tests both demonstrated intercept invariance.

This finding of relatively higher performance for African American students on a simple digit recall task is consistent with previous research comparing racial/ethnic groups on cognitive tasks (e.g., Jensen, 1973; Jensen & Reynolds, 1982). Jensen hypothesized that such differences may be due to a difference, across groups, in Level 1 (simple memory) versus Level II (more complex reasoning) abilities (Jensen, 1973, 1982), and later, that the degree of racial/ethnic group difference is related to the degree of *g*-loading of a test ("Spearman's hypothesis," Jensen, 1998; Reynolds & Jensen, 1983).

Our Cattell-Horn-Carroll-theory-based speculation is that this anomaly is the result of the three tests that make up the Gsm factor in this research measuring different narrow cognitive abilities. As noted elsewhere, a lack of intercept invariance may result from the existence of an unmodeled minor factor (Keith, 2015, Chapter 19). This possibility seems likely with the DAS-II, with two of the tests on the Gsm factor (Recall of Digits Backward and Recall of Sequential Order) measuring working memory skills and one (Recall of Digits Forward) measuring memory span. This possibility could be evaluated by a cross-battery CFA in which one or more additional measures of memory span are included. Of course this possible explanation is not inconsistent with Spearman's hypothesis; a memory span factor also likely has lower loadings on *g* than does a working memory factor (Jensen, 1998, Chapter 8).

## Limitations and Future Research

These results suggest the DAS-II is relatively free of construct bias across the groups in the current study; nevertheless, several limitations need to be considered. First, pertaining to the measure itself, the current study included only those subtests administered to all children in the school-age sample. Further investigation into test bias with the DAS-II should consider the remaining school-age subtests as well as subtests used in the younger age battery to ensure that the evidence for a lack of construct bias also applies to those subtests and age groups. Also, given that there has been investigation into content bias at the item level (Elliott, 2007) and now construct bias (current study and Trundt, 2013), it will be important to examine predictive bias in more depth (beyond the research reported briefly in the DAS-II manual, pp 222-223). In addition, although the ability to assess the reliability of the findings by replicating the initial analyses with three comparison groups is a strength of the current study, the results of this validation process would be strengthened further with additional samples from the racial/ethnic groups studied here, and with samples from additional racial/ethnic groups (e.g., Native American students). Conducting future research that addresses these issues would further bolster evidence for the absence of bias in the DAS-II across groups.

Four broad racial/ethnic groups were included in the current study. Although it is a strength to have an Asian sample in an analysis of test bias (cf. Valencia & Suzuki, 2001, Chapter 5), the sample size for this subsample was still small by CFA standards. If analyzed separately, the Asian sample of 98—although arguably sufficient in power to differentiate a good fit from a poor fit (RMSEA of .05 vs. .10, *df* = 62, power = .87)—is still underpowered by conventional standards (see, for example, Keith, 2015, Chapter 21). The first step in these invariance analyses (configural invariance) is equivalent to analyzing each group separately. Power increases, however, as multigroup constraints are added to the models, and thus the results from the metric and intercept invariance steps for Asian children are likely more stable and accurate than those from the configural invariance step. As noted in Keith (2015), problems caused by small samples size are magnified when there are few indicators of latent variables; most latent variables in these DAS-II models included only two indicators. Taken together, these considerations argue for caution in interpreting the results for Asian children, especially concerning the results from the configural invariance step of the analysis. In particular, the finding of the lack of a coherent Glr factor for Asian children in the configural invariance step may, in fact, be the result of there being no such factor for Asian children, or it may be the result of small sample size and only two indicators of the factor. Additional research with a larger Asian sample, and research with other racial/ethnic groups, is still needed.

### Strengths and Implications

This research also had several strengths. One is the large, nationally representative sample and another is the incorporation of replication samples. These two elements strengthen the generalizability of our findings. In addition, the multigroup CFA methodology allowed for a comprehensive evaluation of differential internal structure validity evidence. Finally, the study incorporated the assessment of intercept invariance/bias. Intercept invariance is important because it assesses whether every subtest has the same zero point across groups; a group with a lower intercept would thus obtain a lower average score on the subtest even when the level of underlying latent ability is the same across groups. Intercept invariance, although important, has been less commonly addressed in construct bias research.

Outcomes of this research support the general appropriateness of the DAS-II for clinical use with several racial/ethnic groups. Psychologists and those who administer and interpret scores from this measure can feel confident that the constructs measured by the DAS-II are likely equivalent for students who are African American, Asian, and Hispanic in comparison to their Caucasian peers. These research findings suggest that the DAS-II is not biased toward or against any of these groups. Testing professionals can also feel reasonably confident that results from the DAS-II, when obtained through a standard administration and considered in conjunction with relevant data from other sources, can provide useful information regarding the cognitive performance of a diverse range of children and youth. As with all cognitive measures, however, it is important to recognize that a single number does not capture the full range of a person's abilities, but rather represents an approximation of that person's performance on a specific task in a certain context.

## Acknowledgments

The authors are grateful to Cindy Carlson, Ricard Valencia, Randy Kamphaus, and Stephanie Cawthon for their contributions to that dissertation. The authors are grateful to Pearson Assessments and Colin Elliott for providing these data, and to Danika Maddocks and Stephanie Young for comments on an earlier version of the manuscript.

## Authors' Note

This article is based, in part, on the first author's doctoral dissertation at the University of Texas, Austin. The authors are responsible for any errors and for the opinions expressed in this article.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## References

Arbuckle, J. L. (2014). IBM SPSS Amos 23 User's Guide [Computer program]. Crawfordville, FL: Amos Development Corporation.

Binet, A., Simon, T., & Kite, E. S. (1916). *The development of intelligence in children (The Binet–Simon Scale)*. Baltimore, MD: Williams & Wilkins. doi:10.1037/11069–000

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling*, *7*, 461-483.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*, 230-258.

Canivez, G. L., & McGill, R. J. (2016). Factor structure of the Differential Ability Scales–Second Edition: Exploratory and hierarchical factor analyses with the core subtests. *Psychological Assessment*, *28*, 1475-1488. doi:10.1037/pas0000279

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255.

Cohen, R. J., & Swerdlik, M. E. (2002). *Psychological testing and assessment: An introduction to tests and measurement* (5th ed.). Mountain View, CA: Mayfield Publishing.

Dumont, R., Willis, J. O., & Elliott, C. D. (2009). *Essentials of DAS-II assessment*. Hoboken, NJ: John Wiley.

Edwards, O. W., & Oakland, T. D. (2006). Factorial invariance of Woodcock-Johnson III scores for African Americans and Caucasian Americans. *Journal of Psychoeducational Assessment*, *24*, 358-366.

Elliott, C. D. (1990). *Differential Ability Scales: Introductory and technical manual*. San Antonio, TX: Psychological Corporation.

Elliott, C. D. (2007). *Differential Ability Scales, Second Edition*. San Antonio, TX: The Psychological Corporation.

Graham, J. W., & Coffman, D. L. (2012). Structural equation modeling with missing data. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 277-295). New York, NY: Guilford Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

IBM Corp. (2014). IBM SPSS Statistics for Windows, Version 23.0 [Computer program]. Armonk, NY: Author.

Jensen, A. R. (1973). Level 1 and level II abilities in three ethnic groups. *American Educational Research Journal*, *10*, 263-276.

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Free Press.

Jensen, A. R. (1982). Level 1/level II: Factors or categories? *Journal of Educational Psychology*, *74*, 868-873.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, *3*, 423-438.

Keith, T. Z. (2015). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*. New York, NY: Routledge.

Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the Differential Ability Scales-II: Consistency across ages 4 to 17. *Psychology in the Schools*, *47*, 676-697. doi:10.1002/pits.20498

Keith, T. Z., Quirk, K. J., Schartzer, C., & Elliott, C. D. (1999). Construct bias in the Differential Ability Scales: Confirmatory and hierarchical factor structure across three ethnic groups. *Journal of Psychoeducational Assessment*, *17*, 249-268.

Keith, T. Z., & Reynolds, M. R. (2012). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 758-799). New York, NY: Guilford Press.

Keith, T. Z., Reynolds, M. R., Roberts, L. G., Winter, A. L., & Austin, C. A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the Differential Ability Scales–Second Edition. *Intelligence*, *39*, 389-404.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting statistical equation analyses. *Psychological Methods*, *7*, 64-82.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525-543.

Reynolds, C. R., & Jensen, A. R. (1983). WISC-R subscale patterns of abilities of Blacks and Whites matched on full scale IQ. *Journal of Educational Psychology*, *75*, 207-214.

Reynolds, C. R., & Lowe, P. A. (2009). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (4th ed., pp. 332-374). New York, NY: John Wiley.

Rushton, J. P., Skuy, M., & Bons, T. A. (2004). Construct validity of Raven's advanced progressive matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment*, *12*, 220-229.

Scheiber, C. (2016a). Do the Kaufman tests of cognitive ability and academic achievement display construct bias across a representative sample of Black, Hispanic, and Caucasian school-age children in grades 1 through 12? *Psychological Assessment*, *28*, 942-952. doi:10.1037/pas0000236

Scheiber, C. (2016b). Is the Cattell–Horn–Carroll-Based Factor Structure of the Wechsler Intelligence Scale for Children–Fifth Edition (WISC-V) construct invariant for a representative sample of African–American, Hispanic, and Caucasian male and female students ages 6 to 16 years? *Journal of Pediatric Neuropsychology*, *2*, 79-88. doi:10.1007/s40817-016-0019-7

Trundt, K. M. (2013). *Construct bias in the Differential Ability Scales, Second Edition (DAS-II): A comparison among African American, Asian, Hispanic, and White ethnic groups* (Doctoral dissertation). Retrieved from https://repositories.lib.utexas.edu/bitstream/handle/2152/21160/TRUNDT-DISSERTATION-2013.pdf

Valencia, R. R., & Suzuki, L. A. (2001). Intelligence testing and minority students: Foundations, performance factors, and assessment issues. Thousand Oaks, CA: Sage.