

DOCUMENT SUMMARY This 2022 study published in JAMA Pediatrics provides statistical evidence that the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)—a "gold standard" assessment—contains racial and gender biases. Despite the authors' conclusion of "minimal bias" overall, the data reveals that 11% of diagnostic items function differently for Black vs. White children and female vs. male children, raising significant concerns about the underdetection of autism in these groups. This research is a critical asset for demonstrating the inherent flaws and discriminatory impact of standardized testing.

FILENAME Kalb_2022_RESEARCH_research_article_ADOS-2_BIAS_racial_sex.md

METADATA Primary Category: RESEARCH **Document Type:** research_article **Relevance:** Core **Update Frequency:** Static **Tags:** #ADOS, #standardized_testing, #assessment_bias, #racial_bias, #gender_bias, #autism, #neurodiversity, #underdetection, #assessment, #eugenics **Related Docs:**

Analysis of Race and Sex Bias in the Autism Diagnostic Observation Schedule (ADOS-2)

Why This Matters to Enliten

This paper is a cornerstone piece of evidence for our mission. It comes from a mainstream, peer-reviewed medical journal (JAMA) and uses advanced statistical methods to prove what we already know: the "gold standard" assessments are biased and fail our clients. Although the authors try to downplay the overall impact by calling the bias "minimal," their own data provides the ammunition we need to dismantle the credibility of the ADOS-2.

We can use these findings to show clients, parents, and other professionals that there is concrete, scientific proof that these tests can lead to the underdetection of autism in Black children and girls. The data on "greater difficulty" and "poorer discrimination" for certain items directly validates our position that these tests don't measure a child's intrinsic neurotype but rather their ability to perform according to a narrow, culturally-biased standard. This paper is a perfect example of the system investigating itself and finding flaws, even if it's unwilling to admit those flaws are catastrophic at the individual level.

Critical Statistics for Our Work

- **Total Sample:** 6,269 children evaluated for ASD.
 - **Race:** 1,619 Black/African American (25.9%), 3,151 White (50.3%).
 - **Sex:** 4,970 male (79.4%), 1,293 female (20.6%).
- **Overall Bias:** 11% of all ADOS-2 diagnostic items analyzed (16 out of 140) showed significant statistical bias (Differential Item Functioning or DIF).
- **Bias by Race (Black/African American vs. White children):**
 - 8 diagnostic items showed significant racial bias.
 - 6 of the 8 biased items (75%) were in the Social Affect (SA) domain.
 - For most biased items, the test had **greater difficulty** and **poorer discrimination** for Black children compared to

- White children. This means a higher level of impairment is needed for a Black child to get the same score as a White child.
 - The authors state this leads to an "**underestimation of ASD severity** for Black/African American children."
 - The item for "repetitive interests" showed a **large** magnitude of bias for Black children.
- **Bias by Sex (Female vs. Male children):**
 - 5 unique diagnostic items showed significant sex-based bias.
 - The "hand mannerisms" item showed bias across **all 3 modules and all 5 algorithms** tested.
 - The magnitude of bias for "hand mannerisms" was **moderate to large**.
 - Given the RRB (Repetitive/Restricted Behaviors) scale only has 4 items, having one of them (**25% of the scale**) consistently underestimate autism in girls is described by the authors as "**notable and worth prompting further research**."
- **ADOS-2 Classification Differences:**
 - A higher percentage of Black children in the sample were classified with "Autism" (60.6%) compared to White children (51.3%).
 - A lower percentage of Black children received a "No ASD/autism" classification (27.4%) compared to White children (36.2%).
 - The mean ADOS-2 Calibrated Severity Score (CSS) was higher for Black children (5.71) than for White children (5.16). (Enliten Note: This suggests Black children may need to present with more severe, obvious symptoms to even get referred for testing, a key systemic bias the paper touches on).

Methodology We Can Learn From

The study used a statistical method called Item Response Theory (IRT) to evaluate Differential Item Functioning (DIF). DIF analysis is a way to statistically determine if a specific test question is harder or performs differently for one group of people versus another, even when both groups have the same underlying level of the trait being measured (in this case, autism severity).

While we reject the psychometric paradigm, understanding the language and methods used to *prove* bias is valuable for our advocacy and for deconstructing the arguments of those who defend these tests. This study provides a template for how bias is statistically identified from within the system itself.

Findings That Challenge the System

Despite the authors' conservative conclusion that the ADOS-2 "does not have widespread systematic measurement bias," their own data provides a powerful indictment of the test.

1. **Bias is Present and Measurable:** The study proves with statistical significance that 11% of the ADOS-2's core diagnostic items are biased across race and sex. This is not a guess; it's a mathematical finding.
2. **Bias Leads to Underdetection:** The consistent pattern found was that biased items were "harder" for Black children and female children. This means the test is structured in a way that is less likely to detect autistic characteristics in these groups, directly contributing to missed or delayed diagnoses. The authors explicitly state, "the DIF

observed for Black/African American and female children could result in underestimation or underdetection of ASD."

3. **Social Items Fail Black Children:** The bias against Black children was most frequently observed in the Social Affect (SA) domain. This supports the argument that the ADOS-2's criteria for "appropriate" social interaction are based on a narrow, White cultural standard, and fails to account for cultural differences in communication styles.
4. **Repetitive Behavior Items Fail Girls:** The bias against girls was most pronounced and consistent in the "hand mannerisms" item. This is critical because it challenges the idea that underdiagnosis in girls is only about social "camouflaging." It shows the test itself is poorly constructed to identify the real-world presentation of female autistic traits. With the RRB scale being so short (only 4 items), having 25% of it be biased is a massive flaw.

Populations Discussed

Black/African American Children

- When bias was present, it was most often in the Social Affect domain.
- The pattern of bias was one of "greater difficulty" and "poorer discrimination," which results in the **underestimation of ASD severity**. The items are less effective at detecting autism in Black children.
- The authors acknowledge that race is a social construct, not biological, and suggest that referral biases may play a role—that Black children may be referred for more general developmental concerns rather than ASD-specific ones, leading to a phenotypically different clinical population. They also cite structural racism and cultural factors related to identifying social affect as potential reasons for these disparities.

Female Children

- The "hand mannerisms" item demonstrated consistent, moderate-to-large bias across all modules.
- This single biased item makes up **25% of the entire RRB diagnostic algorithm**, which the authors concede is "notable."
- The pattern for biased items was generally greater difficulty, which can lead to underestimation of ASD.
- While much of the literature on underdetection in females focuses on social camouflaging, this study provides strong evidence that the issue is also baked into the test's flawed measurement of repetitive behaviors.

Alternative Approaches Mentioned

The paper operates entirely within the standardized testing paradigm. It does not mention or consider clinical interviews, collaborative assessment, or any qualitative, non-standardized methods as a solution. The only proposed solution is more research to potentially tweak the existing biased items, not to question the validity of the entire assessment method. This is a significant gap we can highlight.

Quotes We Might Use

- **On overall findings:** "These findings suggest that the ADOS-2 does not have widespread systematic measurement bias across race or sex. However, the findings raise some concerns around underdetection that warrant further research."
- **On the bottom line:** "At the individual level, the DIF observed for Black/African American and female children could result in underestimation or underdetection of ASD."
- **On bias against Black children:** "...the direction of bias was generally greater difficulty, resulting in underestimation of ASD severity for Black/African American children."
- **On test effectiveness for Black children:** "Discrimination was poorer as well, suggesting these items do not detect ASD as effectively in Black/African American children."
- **On the flawed RRB scale for females:** "Given the brevity of the diagnostic algorithm for RRB, which only includes 4 items, having 25% of the items consistently underestimate ASD in female children is notable and worth prompting further research."
- **On the magnitude of bias:** "The magnitude of DIF was only moderate to large for 2 items: hand mannerisms (among female children) and repetitive interests (among Black/African American children)."

Clinical Implications

The authors state that while the

overall effect on total scores might be small from a statistical/epidemiological perspective, the impact at the *individual level* could be significant. For a specific Black child or female child whose score falls near the diagnostic cutoff, this measurement bias could be the sole reason they are denied a diagnosis and, consequently, access to services and self-understanding. The findings "raise direct concerns about the hand mannerisms item" for females and warrant further research to address the "evidence of underdetection of ASD symptoms in Black/African American children and female children."

Limitations We Can Point To

- The study was conducted at a single specialty clinic, limiting generalizability.
- The study was unable to investigate bias in other racial or ethnic groups (e.g., Hispanic, Asian) due to sample size.
- The authors' conclusion that the bias is "minimal" is a subjective interpretation that downplays the real-world harm of a test that systematically under-detects autism in marginalized groups.
- The entire framework accepts the premise of the ADOS-2 as a valid tool, seeking only to identify statistical anomalies rather than questioning the fundamental flaws of standardized, observational assessment for neurodevelopmental conditions.