Examining Race and Sex Disparities in Wechsler Intelligence Scale for Children: A Differential Item Functioning Analysis with Item Response Theory (IRT)

Angel (Anqi) Zheng[1], George Papandonatos[1]
[1]Brown University School of Public Health, RI

Overview

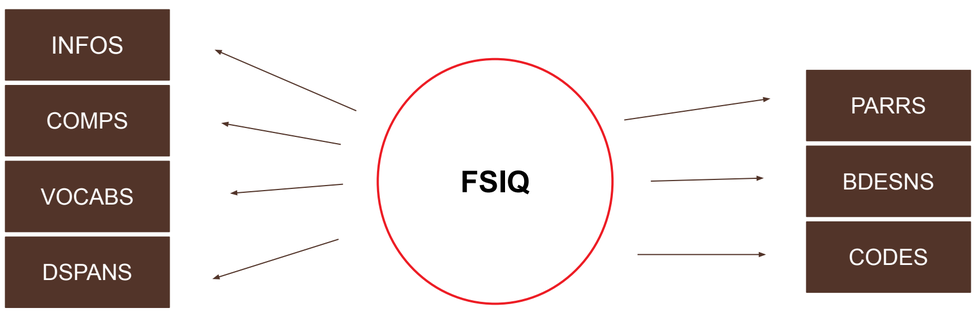
With IQ scores computed based on 7 individual test items from a sample of 3025 subjects, this study examines the disparities between racial groups (Black and White) and sex groups using latent variable modeling packages in R. Each test item is analyzed individually, rather than considering the composite IQ score as a whole.

Background

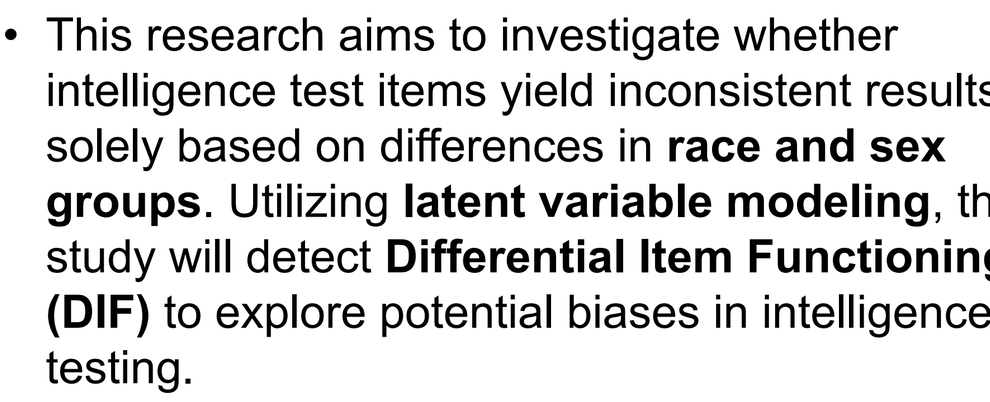
- **The Intelligence Quotient (IQ)** is an abstract, implicit value derived from scores on a series of intelligence tasks. It represents a measure of **latent ability** not directly observed.
- It is crucial that intelligence tasks are unbiased to ensure **fair assessment**
- **The Wechsler Intelligence Scale for Children (WISC)**, developed by psychologist Dr. David Wechsler in 1949, is the most commonly used tool for intelligence measurement.
- This research aims to investigate whether intelligence test items yield inconsistent results solely based on differences in **race and sex groups**. Utilizing **latent variable modeling**, the study will detect **Differential Item Functioning (DIF)** to explore potential biases in intelligence testing.

Method

1 Factor FullScaleIQ Model



Bi-Factor VerbalIQ PerformanceIQ Model



either loaded directly onto FSIQ, or separately onto VIQ and PIQ depending on item type

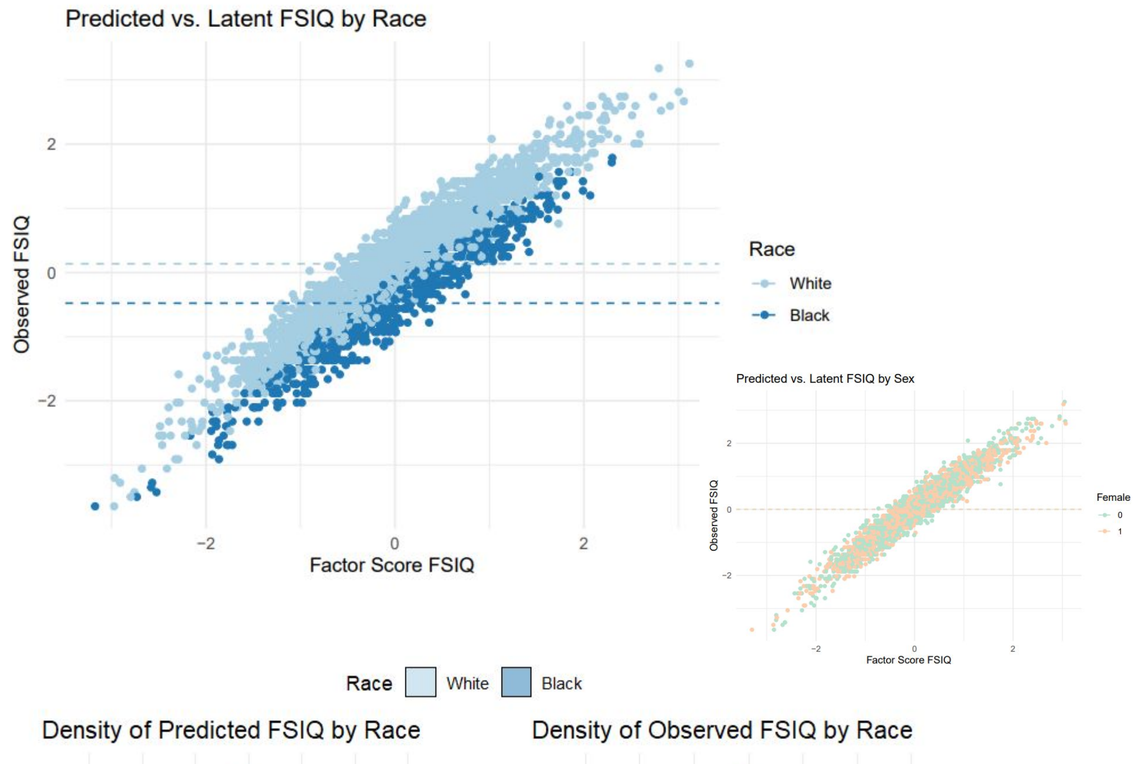
Results

Table 1: Intercepts by Race

	int.W	int.B	Diff	Std Diff	Pval
INFOS	3.0498	3.1973	-0.1474	-1.3730	0.1698
COMPS	3.2020	3.0502	0.1518	1.4439	0.1488
VOCABS	3.0578	2.7630	0.2948	3.0077	0.0026
DSPANS	3.4568	2.8123	0.6445	6.3306	0.0000
PARRS	3.3165	3.1286	0.1879	1.7469	0.0807
BDESNS	3.6934	3.2642	0.4292	3.7972	0.0001
CODES	3.5300	3.2372	0.2928	2.6300	0.0085

Table 2: Loadings by Race

	load.W	load.B	Diff	Std Diff	Pval
INFOS	0.7575	0.6704	0.0871	2.8057	0.0050
COMPS	0.5154	0.4966	0.0188	0.4819	0.6299
VOCABS	0.7343	0.6187	0.1156	3.4980	0.0005
DSPANS	0.5552	0.6320	-0.0768	-2.2373	0.0253
PARRS	0.6295	0.6353	-0.0059	-0.1751	0.8610
BDESNS	0.5139	0.5120	0.0019	0.0493	0.9607
CODES	0.2211	0.3028	-0.0817	-1.7794	0.0752



Conclusion

- Results showed **racial disparity** in WISC 7-item, with some items demonstrating evidence of differential item functioning.
- Vocabulary and digit span show the greatest differential item functioning between race groups.
- Between race groups, there is a **0.692** FSIQ difference in mean, suggesting that the WISC is unfavorably **biased towards White individuals**.
- On the other hand, results **do not show an overall sex disparity**.
- Although some items have different loadings and intercepts between **sex groups**, these differences are observed in both directions and lead to an **overall cancellation** in the total score.

Key References

*Yves Rosseel (2012). **lavaan**: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36. https://doi.org/10.18637/jss.v048.i02*
*R. Philip Chalmers (2012). **mirt**: A Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, 48(6), 1-29. doi:10.18637/jss.v048.i06*