

Journal of Psychoeducational Assessment

<http://jpa.sagepub.com/>

IQ Scores Should Be Corrected for the Flynn Effect in High-Stakes Decisions

Jack M. Fletcher, Karla K. Stuebing and Lisa C. Hughes

Journal of Psychoeducational Assessment 2010 28: 469 originally published online 7 July 2010

DOI: 10.1177/0734282910373341

The online version of this article can be found at:

<http://jpa.sagepub.com/content/28/5/469>

Published by:



<http://www.sagepublications.com>

Additional services and information for *Journal of Psychoeducational Assessment* can be found at:

Email Alerts: <http://jpa.sagepub.com/cgi/alerts>

Subscriptions: <http://jpa.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations: <http://jpa.sagepub.com/content/28/5/469.refs.html>

IQ Scores Should Be Corrected for the Flynn Effect in High-Stakes Decisions

Journal of Psychoeducational Assessment


28(5) 469–473

© 2010 SAGE Publications

Reprints and permission: <http://www.sagepub.com/journalsPermissions.nav>

DOI: 10.1177/0734282910373341

<http://jpa.sagepub.com>



**Jack M. Fletcher¹, Karla K. Stuebing¹,
and Lisa C. Hughes¹**

Abstract

IQ test scores should be corrected for high stakes decisions that employ these assessments, including capital offense cases. If scores are not corrected, then diagnostic standards must change with each generation. Arguments against corrections, based on standards of practice, information present and absent in test manuals, and related issues, ignore expert consensus about the assessment of intellectual disabilities and the acceptance of the Flynn effect in the field. Most psychometric concerns about correction are based on validity studies with small subgroups and do not reflect sufficient effort to estimate the precision of the Flynn estimate. We computed a confidence interval for the Wechsler PIQ across four validity studies that shows a SEM of about 1 around a mean of about 3 points per decade. A meta-analytic weighted mean of the 14 studies in Flynn (2009) is 2.80 (2.50, 3.09), close to Flynn's (2009) unweighted average (2.99). More psychometric research would be helpful, but this level of precision supports the Flynn adjustment of 3 points per decade.

Keywords

IQ, intellectual disability, Flynn effect, Atkins hearings

IQ test scores should be corrected for high-stakes decisions in which a test with older norms is invoked as evidentiary support in the decision-making process. This could include not only Atkins cases involving capital offenses and the death penalty but also intellectual disability (ID) decisions involving social security eligibility or special education where eligibility hinges on a specific score or range of scores. In all these contexts, the person may have previous IQ test scores that are higher than current scores, which may be reconciled by taking into account norms obsolescence.

In Atkins cases as well as other high-stakes assessments, the offender often has multiple IQ scores obtained over a long period of time. Some offenders may have been administered older versions of tests with norms well over 10 years of age, rendering them obsolete and yielding

¹University of Houston, Houston, TX, USA

Corresponding Author:

Jack M. Fletcher, Department of Psychology, University of Houston Texas Medical Center Annex, 2151 W. Holcombe Boulevard, Suite 222, Houston, TX 77204-5053, USA

Email: jackfletcher@uh.edu

inaccurate estimates of IQ (Flynn, 2009). To illustrate, in one case in which the senior author consulted, the offender had WAIS-III (Wechsler Adult Intelligence Scale—Third Edition) scores of 68 and 71, 3 years apart as an adult. As a child, the offender obtained a WISC (Wechsler Intelligence Scale for Children) score of 79 in 1973, 25 years after the normative sample was collected. A correction for the Flynn effect (FE) of 0.3 per year would be $0.3 \times 25 \text{ years} = 7.5$, or 71.5, aligning closely with the WAIS-III assessments. Should an offender be executed because the psychologist who gave the WISC failed to write a note indicating that the IQ score may be an overestimate because of norms obsolescence?

Correcting an IQ score is not a violation of test administration. Rather, it is selecting an appropriate normative comparison (Gresham, 2009). We would not expect pediatricians to use a height/weight chart from another country or century to assess a child's percentile rank in height or weight; if they did, we would expect corrections so that the percentile reflects the current, national distribution. Correcting an IQ score is a simple procedure that avoids having to change standards. Thus, if 15-year-old IQ norms are used, either the score itself must be corrected by about 4.5 points ($0.3 \times 15 \text{ years} = 4.5$) or the cut-point for ID needs to be corrected to 74.5 because the mean IQ of a contemporary sample using the old norms would be 104.5.

Some argue that correcting for norms obsolescence is not a standard of practice (Hagan, Drogin, & Guilmette, 2008; 2010). However, standards of practice are set by consensus reports written by experts. The most prominent guidelines for the assessment of ID represent the 11 editions of the manual for diagnosis by the American Association of Intellectual and Developmental Disabilities (Schalock et al., 2010), not cited by Hagan et al. (2008). Since 2002, this manual has explicitly recommended correcting IQ scores for norms obsolescence, with other researchers agreeing (e.g., Gresham, 2009; Kanaya & Ceci, 2007; Widaman, 2007).

Other objections to correcting for norms obsolescence confuse issues related to why the FE occurs with whether it occurs; its existence is widely accepted, but the cause is disputed (see Flynn, 2010; Kaufman, 2010). There is also confusion involving Flynn's assertion that the WAIS-III norms are problematic (e.g., Flynn, 2009). The publisher's post on this issue (Weiss, 2008) addressed Flynn's claim that there were problems with the norming of the WAIS-III, but has been misinterpreted as indicating that the correction for norms obsolescence was under dispute (Hagan et al., 2008), which is not the case (Zhou, Zhu, & Weiss, 2010). Some suggest that the standardization and validity samples are different and that group data should not be used to correct individual scores (Zhu & Tulskey, 1999). However, individual scores are not being adjusted; rather, the validity studies are used as a basis for selecting an appropriate normative comparison group.

The major questions should involve the magnitude of the effect and its constancy across age and levels of IQ (Tanaka & Ceci, 2007; Zhou et al., 2010). As Widaman (2007) suggested, much of the variation in estimates of the effect is because of measurement error, especially when small samples across different age and IQ levels are used. This variation is important to understand, and it is surprising that more effort has not been expended toward evaluating the precision of the correction.

We estimated 95% confidence intervals (CIs) for the four comparisons of PIQ (Performance IQ) in Zhou et al. (2010) using the standard deviations for each comparison kindly provided by Dr. Xiaobin Zhou (Table 1). The CIs were computed by estimating the standard error of the mean (SEM) of average change and multiplying by ± 1.96 (the critical z value). The SEM for matched pairs is the SD of the difference divided by the square root of N . To create the CI, we used a standard formula [$CI_{.95} = \text{mean difference} \pm z_{.05/2} (\text{SEM})$]. As Table 1 shows, the confidence intervals do not include 0 and extend approximately 1 point (0.1 per year) on either side of the mean difference of about 3 per decade (0.29–0.31 per year). A simple rubric would be 3 ± 1 . An adjustment for Full-Scale IQ (FSIQ) would be similar because it is highly correlated with PIQ. Because the FSIQ is higher in reliability, the CIs may be smaller.

Table 1. Confidence Intervals for PIQ Across Four Wechsler Tests

Tests	Mean Change Per Year	SD Change Per Year	N	Years Between Norming	SE or SD / \sqrt{N}	SE Times 1.96	Lower CI Mean Minus SE \times 1.96	Upper CI Mean Plus SE \times 1.96	Lower CI in Points Per Decade	Upper CI in Points Per Decade
WPPSI-R/III	0.24	0.86	174	13	0.07	0.13	0.11	0.37	1.12	3.68
WISC-III/IV	0.29	0.96	239	12	0.06	0.12	0.17	0.41	1.68	4.12
WAIS-R/III	0.29	0.61	191	16	0.04	0.09	0.20	0.38	2.03	3.77
WAIS-III/IV	0.31	0.81	240	11	0.05	0.10	0.21	0.41	2.08	4.12

Note. PIQ = Performance IQ; SD = standard deviation; SE = standard error; CI = confidence interval; WPPSI-R = Wechsler Preschool and Primary Scale of Intelligence-Revised; WISC = Wechsler Intelligence Scale for Children; WAIS-R = Wechsler Adult Intelligence Scale-Revised.

Table 2. Weighted Mean Effects, Confidence Intervals, and Tests of Homogeneity

Newer Tests	Older Tests	Difference Years	N	Mean Difference	Difference Per Decade	Deviation Squared Model 1	Deviation Squared Model 2	Deviation Squared Model 3
SB-5	WAIS-III	6	87	5.50	9.17	43.61		
SB-4	WAIS-R	7	47	3.42	4.89	5.07	4.29	4.29
WISC-IV	WAIS-III	6.75	198	3.10	4.59	11.85	9.75	
SB-5	WISC-III	12	66	5.00	4.17	1.72	1.33	1.33
WISC-IV	WISC-III	12.75	244	4.23	3.32	1.14	0.53	0.53
WISC-III	WISC-R	17	206	5.30	3.12	0.43	0.10	0.10
SB-4	WISC-R	13	205	2.95	2.27	0.74	1.29	1.29
SB-5	SB-4	16	104	2.77	1.73	2.62	3.50	3.50
WAIS-III	WAIS-R	17	192	4.20	2.47	0.64	1.47	
SB-4	SB-LM	13	139	2.16	1.66	2.09	2.75	2.75
WAIS-R	WISC-R	6	80	0.90	1.50	2.48	3.17	3.17
WAIS-III	WISC-III	6	184	-0.70	-1.17	53.48		0.00
WAIS-IV	WAIS-III	11	240	3.37	3.06	0.64	0.09	0.09
WAIS-IV	WISC-IV	4.25	157	1.20	2.82	0.00	0.05	0.05
Mean effect						2.80	2.96	2.86
Q						126.52***	28.32**	17.10**

Note. SB = Stanford-Binet Intelligence Scale; WAIS = Wechsler Adult Intelligence Scale; WISC; Wechsler Intelligence Scale for Children.

** $p < .003$. *** $p < .0001$.

Table 2 uses the 14 studies in Flynn (2009) to compute the meta-analytic mean, showing an inverse variance weighted mean effect (Lipsey & Wilson, 2001) per decade of 2.80 (2.50, 3.09), close to Flynn’s unweighted average. We tested the distribution of effects for heterogeneity using the Q statistic (which is distributed as a chi-square with $k - 1$ degrees of freedom, where k equals the number of studies), and found that the 14 effects were more variable than would be expected because of sampling error alone, $Q_{(13)} = 126.52, p < .0001$. Although the CI is small, significant heterogeneity potentially limits the usefulness of the mean effect because of averaging dissimilar effects. Inspection of the contribution of each effect to the Q statistic (Deviation Squared Model 1 in Table 1) revealed two outliers, one very large and one very small, both of which involved the WAIS-III. After removing these two outliers, the mean effect per decade was 2.96 (2.65-3.27), with $Q_{(11)} = 28.33, p < .003$. Given the questions raised about the normative sample for the WAIS-III (Flynn, 2009), we removed the other two WAIS-III comparisons and found a mean

effect of 2.86 (2.5-3.22) and $Q_{(9)}=17.1, p < .047$. Thus, the sources of heterogeneity can be identified. We do not view this finding as supporting Flynn's claim that the WAIS-III norms are problematic. Rather, more research with additional samples and perhaps the inclusion of other tests may enhance understanding of factors responsible for the variability across studies and make possible more precise estimates of the effect of norms obsolescence.

These two approaches to estimating the mean and the precision of the effect support Flynn's aggregated estimate of the magnitude of norms obsolescence and are sufficiently precise to justify corrections for high-stakes decisions. There is variability across studies, and age/ability level, but this is true for any subject matter. The estimate of 3 ± 1 is similar to the estimates for the conversion of WAIS-III and WAIS-IV scores for the middle of the distribution (where the sample size is larger) in table 5.6 of the WAIS-IV technical manual.

The administration/technical manuals' silence over the FE has been interpreted in Atkins cases as evidence that scores should not be corrected. Clearly publishers have acknowledged the FE by renorming tests more frequently and providing validity studies and conversion tables. A publisher should not be expected to address every use of the test. The WAIS-IV manual, for example, provides no guidance on the diagnosis of ID. However, Weiss (2008) is commonly invoked as denying that the FE exists (Hagan et al., 2008) when it actually addresses the adequacy of the WAIS-III norms. In one Atkins hearing, an email from the technical assistance hotline of a publisher was introduced in response to a question about the FE from a testifying psychologist. The email indicated that the publisher did not recommend correcting scores. Telephone calls and emails requesting clarification from the publisher elicited no response and the judge cited the email in ruling against the offender.

Publishers may need to do more by providing data like that in Tables 1 and 2 (and studies like Zhou et al., 2010) and by indicating explicitly that when outdated norms are used, corrections will be necessary to appropriately scale the scores. This would facilitate adoption of practices recommended by the American Association of Intellectual and Developmental Disabilities into the different venues where IQ scores are used for high-stakes decision making. IQ scores based on obsolete norms should be corrected and can be estimated with reasonable precision in high-stakes decisions, including capital offense cases. There is no evidence that Flynn's correction overestimates IQ at the lower end of the distribution (Zhou et al., 2010).

Summary and Conclusions

IQ test scores should be corrected for any high-stakes decision that employ these assessments, including capital offense cases. If scores are not corrected, then diagnostic standards must change with each generation. Arguments against correction ignore expert consensus about the assessment of intellectual disabilities and do not take into account the wide acceptance of the FE. More research on the precision of the estimate would be helpful, but the level of precision we reported of a mean of about 3 and a SEM of about 1 supports the correction and is consistent with the Flynn correction of 3 points per decade.

Declaration of Conflicting Interests

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

Funding

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

References

- Flynn, J. R. (2009). The WAIS-III and WAIS-IV: *Daubert* motions favor the certainly false over the approximately true. *Applied Neuropsychology, 16*, 98-104.
- Flynn, J. R. (2010). Problems with IQ gains: The huge vocabulary gap. *Journal of Psychoeducational Assessment, 28*, 412-433.
- Gresham, F. M. (2009). Interpretation of intelligence test scores in *Atkins* cases: Conceptual and psychometric issues. *Applied Neuropsychology, 16*, 91-97.
- Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2008). Adjusting IQ scores for the Flynn effect: Consistent with standard of practice? *Professional Psychology: Research and Practice, 39*, 619-625.
- Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2010). IQ scores should not be adjusted for the Flynn effect in capital punishment cases. *Journal of Psychoeducational Assessment, 28*, 474-476.
- Kanaya, T., & Ceci, S. J. (2007). Mental retardation diagnosis and the Flynn effect: General intelligence, adaptive behavior, and context. *Child Development Perspectives, 1*, 62-63.
- Kaufman, A. S. (2010). "In what way are apples and oranges alike?": A critique of Flynn's interpretation of the Flynn effect. *Journal of Psychoeducational Assessment, 28*, 382-398.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Schallock, R., Borthwick-Duffy, S., Bradley, V., Buntinx, W., Coulter, D., Craig, E., . . . Yeager, M. (2010). *Intellectual disability: Definition, classification, and systems of support* (11th ed.). Washington, DC: American Association on Intellectual and Developmental Disabilities.
- Weiss, L. G. (2008). *WAIS-III technical report: Response to Flynn*. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/98BBF5D2-F0E8-4DF6-87E2-51D0CD6EE98C/0/WAISIII_TR_lr.pdf
- Widaman, K. (2007). Stalking the roving IQ score cutoff: A commentary on Kanaya and Ceci. *Child Development Perspectives, 1*, 57-59.
- Zhou, X., Zhu, J., & Weiss, L. G. (2010). Peeking inside the "black box" of the Flynn effect: Evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment, 28*, 399-411.
- Zhu, J., & Tulskey, D. S. (1999). Can IQ gain be accurately quantified by a simple difference formula? *Perceptual and Motor Skills, 88*, 1255-1260.