JARID Journal of Applied Research in Intellectual Disabilities
bild Published for the British Institute of Learning Disabilities
Journal of Applied Research in Intellectual Disabilities 2014

Intelligence is in the Eye of the Beholder: Investigating Repeated IQ Measurements in Forensic Psychiatry

Petra Habets*, Inge Jeandarme*, Kasia Uzieblo[†,‡], Karel Oei[§] and Stefan Bogaerts[¶,**]

*Knowledge Centre Forensic Psychiatric Care (KeFor) OPZC Rekem, Rekem, Belgium; [†]Thomas More, University College, Antwerp, Belgium; [‡]Catholic University, Leuven, Belgium; [§]Tilburg University, Tilburg, The Netherlands; [¶]School of Social and Behavioral Sciences, Developmental and Forensic Psychology, Tilburg, The Netherlands; **KARID, Forensic Psychiatric Centre The Kijvelanden, Rotterdam, The Netherlands

Accepted for publication *10 July 2014*

Background A stable assessment of cognition is of paramount importance for forensic psychiatric patients (FPP). The purpose of this study was to compare repeated measures of IQ scores in FPPs with and without intellectual disability.
Methods Repeated measurements of IQ scores in FPPs ($n = 176$) were collected. Differences between tests were computed, and each IQ score was categorized. Additionally, *t*-tests and regression analyses were performed.
Results Differences of 10 points or more were found in 66% of the cases comparing WAIS-III with RAVEN scores. Fisher's exact test revealed differences between two WAIS-III scores and the WAIS categories. The WAIS-III did not predict other IQs (WAIS or RAVEN) in participants with intellectual disability.
Discussion This study showed that stability or interchangeability of scores is lacking, especially in individuals with intellectual disability. Caution in interpreting IQ scores is therefore recommended, and the use of the unitary concept of IQ should be discouraged.

Keywords: cognitive ability, Groninger Intelligence Test, intelligence tests, IQ, psychometrics, RAVEN, repeated measures, stability, WAIS

Introduction

A stable assessment of cognitive functioning (i.e. intelligence) is of paramount importance for forensic psychiatric patients. The level of an individual's intellect has an impact on interrogations, court proceedings, court rulings, risk assessments and treatment programs. Consequently, countries have specific procedures regarding offenders with intellectual disability (OIDs). For example, in Belgium and the Netherlands, if an individual who committed a crime has a diagnosis of intellectual disability, it is possible that he/she will not be held responsible for his/her actions (not guilty by reason of insanity). As a result, a protection measure will be ordered (van Emmerik 2001; Verlinden *et al.* 2009). Additionally, in most states of the United States, people with OIDs are not allowed to be executed. The assessment of intellectual disability can therefore literally be a matter of life and death, leaving no room for error (Fabian *et al.* 2011). Despite these concerns, uniformity is still lacking in assessment of intelligence in forensic populations. Different tools that do not – or only partly – measure the same aspects of intelligence are used, resulting in poorly interchangeable scores (McBrien 2003; Uzieblo *et al.* 2012). It is therefore critical that intelligence is measured in a valid and stable manner and composite scores should be avoided. Namely, it is widely acknowledged that intelligence has a hierarchical structure (i.e. the Cattell–Horn–Carroll model) (McGrew 2009), and minimizing intelligence into a single score fails to captivate the complexity of a person's intellect especially in persons with borderline intelligence (Uzieblo *et al.* 2012).

10.1111/jar.12120

Although large correlations between IQ tests have been reported, research has shown that scores obtained on intelligence tests given to the same individual are not identical (Floyd *et al.* 2008; Di Nuovo *et al.* 2012). In fact, IQ scores are not expected to have perfect instrumental or temporal stability (Evans 1991). Studies regarding stability and consistency between and within IQ tests have shown positive results. For example, Wechsler (1997) reported a 0.91 stability coefficient (i.e. the correlation between assessments using the same test within the same individual) of the Wechsler Adult Intelligence Scale III (WAIS-III) with 1 month separating the two assessments (Wechsler 1997). However, these coefficients were acquired in individuals within the normal IQ range within a short time period, making it not necessarily representative for individuals with intellectual disability. A meta-analysis by Whitaker (2008b) investigated stability coefficients in individuals with intellectual disability and found reasonable stability for full-scale IQs (0.82). Despite the relative stability of scores, a 10-point change or more between assessments with the same instrument was found in 14% of the subjects. Investigating differences in IQ between instruments in an intellectual disability population, Silverman *et al.* (2010) found a mean difference between the WAIS (Wechsler 1955) and Stanford–Binet (Roid 2003) scores of 16.7 points in which the WAIS scored systematically higher than Stanford–Binet. A difference of 10 points or more was found in 85% of the individuals when comparing tests, and 24% had a 20-point difference or more. In contrast, they reported a strong correlation between the two tests ($r = 0.82$) indicating that, despite the large differences between the two instruments, they measured the same basic construct (Silverman *et al.* 2010). Research investigating stability of IQ scores within and between instruments in intellectual disability is scarce and even more so in forensic psychiatric populations. A recent Dutch study investigated the stability of IQ scores in a forensic psychiatric sample. IQ measurements – WAIS-III, Groninger Intelligence Test (GIT; Kooreman & Luteijn 1987) and Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman 1993) – were collected and compared for 50 individuals. They found that when using the WAIS-III to determine intellectual disability, only eight individuals fell into the intellectual disability category, whereas when using the GIT and KAIT, 17 and 29 individuals, respectively, fell into the intellectual disability category. Additionally, about half of the individuals had a difference of at least 10 points when comparing the KAIT with the WAIS-III and the KAIT with the GIT (Van Toorn & Bon 2011).

In sum, research has shown reasonable stability coefficients within tests and relatively high correlations between tests. However, large differences in IQ scores within individuals are possible, which consequently can have severe implications. Furthermore, the question remains whether studies of stability and interchangeability of IQ scores can be translated to the intellectual disability population. Recent evidence suggests that this might not be the case. The purpose of this study was to describe and compare repeated measurements of intelligence in a forensic psychiatric sample with and without intellectual disability. It was predicted that different IQ tests would result in different classifications of intellectual disability. Consequently, a different pattern of regression coefficients was expected to be found in individuals with intellectual disability when compared to individuals without intellectual disability.

Methods

Sample and participant selection

This study is part of a large observational study, which is the first study in Flanders investigating recidivism in forensic psychiatric patients. Patients who were admitted between 2001 and 2010 to one of the three medium security forensic wards in Bierbeek, Rekem or Zelzate ($n = 542$) were eligible to be included in the study. Eleven patients refused participation, resulting in a final sample of 531 participants. Data were gathered by accessing prison and psychiatric hospital records. Information regarding level of education, psychiatric diagnosis, criminal history, hospitalization/imprisonment periods and IQ scores was collected. Diagnosis was based on the Diagnostic and Statistical Manual of Mental Disorder-IV text revision (DSM-IV-TR; American Psychiatric Association 2000).

Assessments and measures

The following intelligence tests were found: the Dutch adaptation of the Wechsler Adult Intelligence Scale (WAIS; Wechsler 1955, 1970), the WAIS-III (Wechsler 1997, 2005), Raven's Progressive Matrices (RAVEN; Raven *et al.* 1998) and the short Groninger Intelligence Test (sGIT; Kooreman & Luteijn 1987). Of the 531 participants, 176 (33%) had two or more IQ scores. The place of administration of the IQ tests is presented in Table 1. Reports of IQ tests can come from psychiatric centres, penitentiaries and from forensic psychiatric assessments (FPA's). A FPA is ordered by a judge to

assess whether or not the offender is accountable for his or hers crimes and can include results of IQ tests. This assessment can take place in the penitentiary, in a psychiatric centre or in an ambulatory setting. The mean age at the time of administration of the IQ test and the corresponding sample size are reported in Table 2.

WAIS and WAIS-III

The WAIS measures general intelligence or 'g' and is divided into two parts: the verbal scale and the performance scale. Each of these two parts is further divided into subtests, each of which taps a specific verbal or non-verbal skill (Wechsler 1955). The WAIS-III is the revised version of the WAIS-R (the successor of the WAIS). Given that the WAIS-R has never been translated into Dutch, no scores are available for the WAIS-R. Two different Belgian norms from 2000 to 2005 are available for the WAIS-III (Wechsler 2000, 2005; Tellegen 2002). However, the norm table that was used was not found in the the most recent majority of reports, making it impossible to recalculate the full-scale WAIS-III scores using norms.

Table 1 Number of IQ tests stratified for place of IQ test administration

	WAIS-III		*WAIS*		*RAVEN*		*sGIT*	
	# Obs.	*%*	*# Obs.*	*%*	*# Obs.*	*%*	*# Obs.*	*%*
Psychiatric centre	125	75	30	49	9	12	0	
Penitentiary	12	7	11	18	3	4	0	
FPA	3	2	12	20	60	78	31	97
Other	2	1	4	7	0	0	0	
Unknown	24	14	4	7	5	6	1	3
Total	166		61		77		32	

Only participants with more than one IQ score on record are included in this table.
FPA, Forensic psychiatric assessment; # Obs., number of observations per test (not equal to number of subjects as subjects have more than one IQ test).

Table 2 Differences in age at time of testing

	Mean (SD)	*Mean difference (SD)*	*t*	*d.f.*	*n*
WAIS-III$^{(1)}$	35.02 (9.57)	−1.42 (3.38)	−2.75	42	43
WAIS-III$^{(2)}$	36.44 (9.00)				
WAIS-III	40.24 (8.86)	8.63 (5.43)*	11.69	53	54
WAIS	31.61 (9.43)				
WAIS-III	34.86 (8.17)	3.68 (3.99)*	7.75	70	71
RAVEN	31.18 (8.87)				
WAIS-III	31.41 (8.33)	1.34 (2.54)*	2.85	28	29
sGIT	30.07 (9.50)				

*Significant at $P < 0.01$.
SD, standard deviation; *n*, number of observations where age at the moment of testing was available.

RAVEN

The RAVEN is a non-verbal intelligence test that requires inductive reasoning about perceptual patterns and is considered to be a good measure for g and more specifically, 'fluid' g (Tulkin & Newbrough 1968; Schroth 1983). Moreover, it has been shown to be a valid instrument in cross-cultural research (Jensen 1980; Raven *et al.* 1983). Given that in Belgium many different norms are available (Moenaert 2006), RAVEN raw scores were transformed using the latest Belgian norms (Magez *et al.* 2006).

sGIT

The short version of the GIT2 (sGIT) (Luteijn & Barelds 2004) consists of six subtests (the full version contains 10 subtests) and is comparable to the WAIS. Studies have found a correlation of $r = 0.94$ between the sGIT and the GIT2, concluding that the sGIT can be administered without problems.

Statistical analyses

The WAIS-III was used as the reference score because it was the most frequently available score among participants who had more than one IQ score. When two WAIS-III scores are reported, the lowest score found in the database will be presented as WAIS-III$^{(1)}$ and the other WAIS-III score as WAIS-III$^{(2)}$. Difference scores were computed by subtracting the corresponding second IQ score from the WAIS-III score within a subject (WAIS-III$^{(1)}$ – WAIS-III$^{(2)}$; WAIS-III – WAIS; WAIS-III – RAVEN; WAIS-III – sGIT). Frequencies of the absolute difference scores are presented in Figure 1. Paired sample *t*-tests were performed to investigate whether WAIS-III$^{(1)}$ scores significantly differed from WAIS-III$^{(2)}$, WAIS, RAVEN or sGIT scores. For each IQ test, IQ scores were divided into categories: 1 = normal IQ (≥85), 2 = borderline IQ (71–84) and 3 = intellectual disability (≤70). Categorical differences between IQ scores were tested using Fisher's exact test. To investigate whether one IQ score predicted another IQ score,

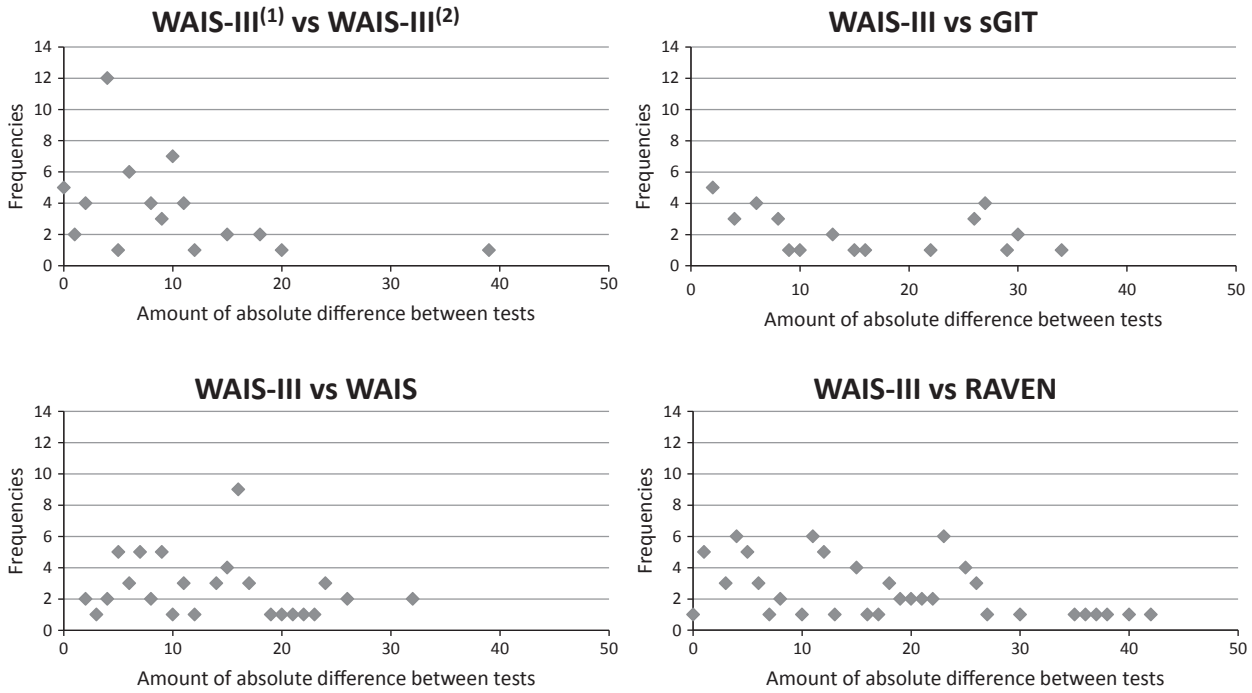

Figure 1 Frequencies of difference scores. Frequencies of absolute differences between tests are reported on the *y*-axis, and on the *x*-axis, the amount of absolute difference between tests is reported. For example, for the WAIS-III$^{(1)}$ versus WAIS-III$^{(2)}$, one person has a difference of 39 points, whereas 12 persons have a difference of four points and five persons have no differences between scores.

multilevel regression analyses were conducted using the XTREG command in STATA (StataCorp 2011) because of the two-level grouping structure of the data, compromising statistical independence of the observations, namely IQ scores (level 1) were nested in subjects (level 2). WAIS-III$^{(1)}$ score was used as the independent variable, WAIS-III$^{(2)}$ score, WAIS score, RAVEN score or sGIT score as the dependent variable, and subject number was modelled as random effect. Multilevel regression analyses were repeated with the year of IQ test administration as a covariate. To examine whether the level of association between IQ scores differed by education level or diagnosis of intellectual disability, multilevel regression analyses were repeated stratified by education level (1 = normal education, 2 = special needs education) and diagnosis of intellectual disability (1 = no diagnosis of intellectual disability, 2 = diagnosis of intellectual disability).

Results

Demographic characteristics

Of the 167 participants, 5 (3%) were female. Participants exhibited the following Axis I diagnoses: developmental disorders (6%, n = 13), substance-related disorders (46%, n = 103), psychotic disorders (18%, n = 41), mood disorders (6%, n = 6), panic disorders (1%, n = 2), paraphilia (5%, n = 11), cognitive disorders (1%, n = 1), other disorders (17%, n = 37) and no or postponed diagnosis (4%, n = 9). Axis II diagnoses established in the participants were cluster A personality disorders (7%, n = 14), cluster B personality disorders (45%, n = 94), cluster C personality disorders (5%, n = 10), personality disorders NOS (13%, n = 28), intellectual disability (21%, n = 44) and no or postponed diagnosis (10%, n = 21). In total, 44 participants had a diagnosis of intellectual disability and 43 participants had been enrolled in special needs education (and four participants had missing values for education). Age at the time of testing differ significantly for all comparisons such that WAIS-III scores were from older individuals compared to the other tests (Table 2). In a number of cases, the amount of time between assessments was <1 year (WAIS-III$^{(1)}$ – WAIS-III$^{(2)}$ = 26%, WAIS-III – RAVEN = 25%, WAIS-III – sGIT = 24%).

Difference scores

When comparing the two WAIS-III scores, 33% of the cases had difference scores that were higher than 10 points. For

the comparisons of the WAIS-III with WAIS, RAVEN and sGIT scores, 60, 66 and 52% of the cases, respectively, had difference scores higher than 10 points (Figure 1).

Differences between scores

All comparisons testing differences between IQ scores were significant (Table 3). The largest mean difference was found between the WAIS-III full-scale score and the sGIT (13.49), and the smallest mean difference was between the verbal IQ (VIQ) score on the WAIS-III$^{(1)}$ and the VIQ score on the WAIS-III$^{(2)}$ (−5.25).

Change in category

Cross-tabulation analyses using Fisher's exact test revealed significant differences in IQ categories between the WAIS-III$^{(1)}$ and the WAIS-III$^{(2)}$ and the WAIS-III and WAIS (Table 4). When comparing the WAIS-III$^{(1)}$ categories with the WAIS-III$^{(2)}$ categories, 15 of the 55 (27%) cases changed category: five (9%) from borderline to normal, nine (16%) from intellectual disability to borderline and one (2%) from intellectual disability to normal. For the WAIS-III/WAIS comparison, 30 of the 62 cases (48%) changed category: 20 (32%) from borderline to normal, six (10%) from intellectual disability to borderline and three (5%) from intellectual disability to normal. The WAIS-III/RAVEN and WAIS-III/sGIT categorical difference comparisons reached trend significance (Table 4). For the WAIS-III/RAVEN comparison, 47 of the 77 cases (61%) changed category: 17 (22%) from borderline to normal, 19 (25%) from intellectual disability to borderline and 11 (14%) from intellectual disability to normal. For the WAIS-III/sGIT comparison, 18 of the 33 cases (55%) changed category: 10 (30%) from borderline to normal, one (3%) from intellectual disability to borderline and seven (21%) from intellectual disability to normal. Changes in category were not associated with time of administration, for example changes from normal to intellectual disability were not associated with longer duration between tests (results available upon request).

Regression analyses

The WAIS-III$^{(1)}$ IQ scores significantly predicted the WAIS-III$^{(2)}$, WAIS and RAVEN IQ scores (P's < 0.001). Adding year of administration as a covariate to the model did not significantly change the direction of effect nor the P-values. Stratified analyses revealed that among participants with a history of special needs education or a diagnosis of intellectual disability, WAIS-III IQ scores did not significantly predict WAIS or RAVEN IQ scores (Table 5).

Discussion

The stability and/or exchangeability of IQ scores was investigated in a forensic psychiatric sample with and

Table 3 Differences and correlations between IQ scores

	Mean (SD)	*Mean difference (SD)*	*t*	*d.f.*	*r*
WAIS-III$^{(1)}$FSIQ	75.00 (14.67)	−7.44 (6.46)*	−8.53	54	0.90*
WAIS-III$^{(2)}$FSIQ	82.44 (14.15)				
WAIS-III$^{(1)}$VIQ	74.60 (17.06)	−5.25 (5.00)*	−6.88	42	0.96*
WAIS-III$^{(2)}$VIQ	79.86 (16.94)				
WAIS-III$^{(1)}$PIQ	75.39 (13.27)	−7.21 (6.22)*	−7.60	42	0.89*
WAIS-III$^{(2)}$PIQ	82.60 (13.33)				
WAIS-III-FSIQ	83.36 (17.01)	−8.54 (12.44)*	−5.38	60	0.74*
WAIS-FSIQ	91.90 (17.10)				
WAIS-III-VIQ	81.56 (14.21)	−8.83 (8.49)*	−7.49	51	0.84*
WAIS-VIQ	90.38 (15.58)				
WAIS-III-PIQ	80.98 (13.15)	−12.62 (13.55)*	−6.72	51	0.60*
WAIS-PIQ	93.60 (16.35)				
WAIS-III-FSIQ	83.65 (17.46)	10.58 (15.15)*	6.13	76	0.54*
RAVEN	73.06 (13.11)				
WAIS-III-FSIQ	79.61 (19.23)	13.49 (12.03)*	−6.44	32	0.79*
sGIT	93.09 (17.05)				

*Significant at $P < 0.001$.
SD, standard deviation; *r*, correlation coefficient.

Table 4 Cross-tabulation categories of IQ

	WAIS-III[2]			*WAIS*			*RAVEN*			*sGIT*		
WAIS-III[1]	*≥85*	*71–84*	*≤70*	*≥85*	*71–84*	*≤70*	*≥85*	*71–84*	*≤70*	*≥85*	*71–84*	*≤70*
≥85	14	0	0	22	3	0	10	14	11	7	0	0
71–84	5	18	0	17	6	3	3	11	14	10	7	0
≤70	1	9	8	3	3	4	0	5	9	7	1	1
Fisher's exact test	0.00			0.00			0.07			0.07		

Fisher's exact test: *P*-value's are reported.

without intellectual disability. The results showed high correlations between tests and IQ scores on one test significantly predicted scores on the other IQ tests, suggesting good stability between scores. However, when looking separately at individuals with a diagnosis of intellectual disability or history of special needs education, the stability between scores disappeared. In these individuals, a significant association between two IQ scores was only established when comparing the two WAIS-III scores. Furthermore, all comparisons between tests revealed significant differences between scores, with mean absolute differences larger than 10 points when comparing WAIS-III full-scale IQ and performance scale with the WAIS full-scale IQ and performance scale and when comparing the WAIS-III with the RAVEN and sGIT. Frequencies of difference scores between tests also show substantial dissimilarities between tests, with a percentage of cases having more than a 10-point difference between tests ranging from 33% to 66%. These percentages are comparable to a study investigating IQ stability in a forensic psychiatric population (Van Toorn & Bon 2011). In contrast, these percentages are much higher than those reported in the meta-analysis by Whitaker (2008b) investigating stability coefficients in individuals with low IQ (14%). The range of differences between scores is surprising, but the fact that there are differences is not. As mentioned in the introduction, IQ scores are not expected to have perfect temporal and instrumental stability, and there are several possible explanations for the differences in IQ scores. Certain factors such as dietary changes (Bellisle 2004; Koyama *et al.* 2012; Smithers *et al.* 2012) and changes in quality of education or intellectual stimulation can result in a true change of IQ.

Chance error

All psychometric instruments are influenced by error and such is the case with the assessment of intelligence. Several sources of error are possible which can be classified in two broad categories: chance error and systematic error (Whitaker 2010). Examples of chance error are: fluctuations in test performance or examiner's behaviour, cooperation of the test subject and other personal and environmental factors. In a forensic psychiatric population, cooperation of the test subject can likely have a larger effect than expected in a non-forensic psychiatric population. It is possible that an individual intentionally performs worse to avoid prison (malingering) or is not motivated enough during the assessment as a result of his/hers psychiatric profile. Furthermore, the added stress of being arrested and sent to prison can result in lower scores. For example, Biles (1968) found significantly lower IQ scores upon arrival in prison compared to IQ scores obtained at a later time point during imprisonment. However, other research, albeit with a different approach, has found no effect of long-term imprisonment on IQ (Banister *et al.* 1973; Bolton *et al.* 1976; Goethals 1981; Dettbarn 2012). Administration of an IQ test within a prison setting can also influence test scores. For example, obstacles such as a lack of privacy and adequate space, scheduling conflicts and noise pollution could have an impact on test scores.

Systematic error: Flynn effect

An example of systematic error is the Flynn effect. The Flynn effect refers to the observation (Flynn 1984) that every restandardization sample for a major intelligence test resulted in an IQ score increase of approximately 0.33 points per year. The Flynn effect seems most prominent in people at the lower end of the distribution and in RAVEN scores (Teasdale & Owen 1989; Colom *et al.* 2005; Williams 2013). For example, Teasdale & Owen (2005) found that the Flynn effect primarily reduced the number of low-end scores, resulting in an

Table 5 Multilevel regression analyses

	B	*P*	*95% CI*	*n*
WAIS-III[(1)]FSIQ				
WAIS-III[(2)]FSIQ	0.92	0.00	0.78 to −1.07	55
Normal education	0.86	0.00	0.64 to −1.08	35
Special needs education	0.87	0.00	0.65 to −1.09	20
No diagnosis of intellectual disability	0.85	0.00	0.65 to −1.05	40
Diagnosis of intellectual disability	0.68	0.00	0.35 to −1.01	15
WAIS FSIQ	0.79	0.00	0.60 to −0.98	61
Normal education	0.59	0.00	0.39 to −0.79	51
Special needs education	0.03	0.95	−0.80 to −0.86	7
No diagnosis of intellectual disability	0.58	0.00	0.38 to −0.79	48
Diagnosis of intellectual disability	0.21	0.51	−0.41 to −0.83	13
RAVEN	0.59	0.00	0.33 to −0.84	77
Normal education	0.57	0.00	0.25 to −0.90	53
Special needs education	0.36	0.15	−0.13 to −0.84	22
No diagnosis of intellectual disability	0.53	0.00	0.23 to −0.82	56
Diagnosis of intellectual disability	0.01	0.97	−0.56 to −0.54	21
sGIT	0.07	0.28	−0.06 to −0.19	33
Normal education	0.41	0.00	0.33 to −0.49	22
Special needs education	0.38	0.02	0.5 to −0.71	11
No diagnosis of intellectual disability	0.03	0.58	−0.07 to −0.14	24
Diagnosis of intellectual disability	−0.08	0.80	−0.72 to −0.55	9
WAIS-III[(1)]VIQ				
WAIS-III[(2)]VIQ	0.97	0.00	0.87 to −1.07	43
Normal education	0.92	0.00	0.80 to −1.05	28
Special needs education	1.06	0.00	0.83 to −1.30	15
No diagnosis of intellectual disability	0.93	0.00	0.80 to −1.05	31
Diagnosis of intellectual disability	0.84	0.00	0.45 to −1.22	12
WAIS-III[(1)]PIQ				
WAIS-III[(2)]PIQ	0.89	0.00	0.74 to −1.05	43
Normal education	0.98	0.00	0.82 to −1.15	28
Special needs education	0.57	0.00	0.30 to −0.83	15
No diagnosis of intellectual disability	0.96	0.00	0.78 to −1.14	31
Diagnosis of intellectual disability	0.60	0.00	0.32 to −0.87	12

B, regression coefficients from multilevel regression analyses; *P*, *P*-value; CI, confidence interval; *n*, sample size.

increased number of moderately high scores, with no increase in very high scores. In contrast, some studies have found a reverse Flynn effect with declining scores for those pursuing higher academic education and those not doing so (Teasdale & Owen 2008; Dutton & Lynn 2013).

In the present study, correction for the Flynn effect for the WAIS-III was not possible due to lack of information concerning which norms were used to calculate the WAIS-III scores. RAVEN scores from the present study were transformed using the latest Belgian norms available (Magez *et al.* 2006), thereby reducing potential increases as a result of the Flynn effect. In addition, due to the random sampling of scores, time of administration is balanced between subjects, again minimizing the potential impact of the Flynn effect in these analyses. Furthermore, although the increases found in this study are much larger than would be expected as a result of the Flynn effect alone, the Flynn effect could explain some of the differences in scores. Flynn stated that differences in scores over time do not reflect changes in that person's true IQ score, rather the differences are a result norms change (Flynn 2006).

In contrast, researchers have also stated that because the Flynn effect concerns a rise in average IQ when comparing generations, it does not apply to within-subject test–retest reliability (Rodgers 1998; van Winkel *et al.* 2006).

The current study investigated differences between scores when the same individual is given the same test (WAIS-III[(1)] versus WAIS-III[(2)]) and if the same individual is given different tests (e.g. WAIS-III versus RAVEN). Therefore, it should be noted that differences in scores can have different causes in the former than in the latter. When comparing scores within the same instrument, changes are mainly due to chance error (Whitaker 2010), whereas when comparing two instruments, both chance error and systematic error could result in changes in scores. Other examples of systematic error, that is floor effect and differences between IQ scales, are discussed in detail in Whitaker (2010).

Change due to mental disorder

In psychotic disorder, there is much debate about a potential progressive decline in cognitive functioning (Zampera 1999; Heaton *et al.* 2001). In a 10-year follow-up study, pre-morbid IQ and post-morbid WAIS scores were compared in first episode patients with psychotic disorder. The results showed that patients with high pre-morbid IQs (≥108) had a 10-point decline in cognitive functioning; however, a restoration to pre-morbid level

was found at follow-up (an average of 10 years later). In the low pre-morbid IQ group, a stable course of IQ was found (van Winkel *et al.* 2006). In the current sample, 18% of the participants had a diagnosis of psychotic disorder, which could explain some of the differences between scores. Unfortunately, it was not possible to investigate differences in stability of IQ scores between diagnoses in the present study due to a lack of power. Potential fluctuations of IQ due to a specific mental disorder should be taken into account when interpreting test results of a forensic psychiatric patient.

Thus, variation in scores may or may not represent the individual's true level of intellectual functioning. The term standard error of measurement is used to capture this variability and to provide a statistical confidence interval (CI) within which it is expected the individual's true score falls. Therefore, it is considered good practice to report CIs together with the full-scale IQ score. Most IQ tests report CIs of approximately 10 points (i.e. five points below or above the true IQ) (Whitaker 2008a). For example, an individual's score of 70 on the WAIS-III corresponds with a 95% CI of 67–75 (Wechsler 2005). The present study found absolute differences of more than 10 points in 18 of the 55 cases when comparing two WAIS-III scores and in 51 of the 77 cases when comparing the WAIS-III with the RAVEN, without even taking into account the level of education or diagnosis of intellectual disability. Therefore, depending on the test used, in 33% of the cases or even in 66% of the cases, the person's second score did not fall within his or her reported CI for the first score. This raises some implications for the interpretation of CIs in psychological reports.

Stability of IQ

In the previous sections, several explanations are given for the differences found in this study. However, these explanations do not alter the fact that disparity between test scores needs to be kept as minimal as possible, especially given the large consequences of inconsistent assessments of cognitive abilities for a forensic psychiatric patient. Intelligence does seem to be fairly stable across the lifespan (Deary *et al.* 2004; Gow *et al.* 2011). Deary *et al.* (2004) investigated old intelligence scores from a sample of 90 000 Scottish children at ages 10 and 11 and reassessed them at the age of 80. They found a positive correlation of 0.66 between the two scores. However, this level of stability cannot simply be presumed in individuals with intellectual disability. Silverman *et al.* (2010) compared Stanford–Binet scores with WAIS scores in 74 individuals with intellectual disability. They found that when using the Stanford–Binet scores, 95% of their sample met the criteria for benefits through the Social Security Administration. In contrast, when using the WAIS scores, only 61% of the participants met the same criteria, resulting in a large number of individuals failing to comply with the criteria although their diagnosis of intellectual disability was already established and documented. Similar results are found in the present study. Depending on the type of test used, some individuals are classified as having an intellectual disability or are considered to have a normal IQ.

The American Association on Intellectual and Developmental Disabilities (AAIDD) postulates that a psychometric instrument performs best when used with individuals who score within 2–3 standard deviations of the mean (Schalock *et al.* 2010, p. 39), whereas Whitaker (2013) states that IQ test perform reasonably well within one SD. Individuals with the diagnosis of intellectual disability fall in the extreme left tail of the IQ distribution (e.g. 2–3 standard deviations below the mean). It is therefore not surprising that the associations between IQ scores found in the present study disappeared when the present authors stratified on the basis of intellectual disability diagnosis and educational level. Nevertheless, the consequences of unstable IQ measurements can be great. As a rule, false positive diagnoses are expected to be rare because intellectual disability should not be diagnosed solely on the basis of IQ score. Three criteria need to be met before diagnosing intellectual disability: (i) significant limitations in intellectual functioning, (ii) significant limitations in adaptive behaviour and (iii) age of onset before the age of 18 (American Psychiatric Association 2000). In addition, an IQ score should not be viewed in isolation but should always be interpreted using environmental context, educational history and functioning of adaptive behaviour. However, no safeguard exists for a false negative. The problems faced by individuals who have intellectual disability but do not receive the diagnosis of intellectual disability can be significant, and the risk of a missed diagnosis is even higher in the people who fall within the borderline category. The current study showed significant differences in IQ categories when using different IQ tests and even when using the same test at different times. Implementing an unified model of cognitive abilities in diagnostics could aid in avoiding false negatives or incorrect diagnoses and would help finding a better alignment between treatment and disabilities. A widely cited unified model on cognitive abilities is the Cattell–Horn–Carroll (CHC) model. This

model is an empirically based model that approaches cognition as a multifactorial concept and is regarded as one of the most well-validated hierarchical taxonomies to classify and describe human cognitive abilities (McGrew 2009). The CHC model could help to better disentangle learning disabilities, language disorders and intellectual disability and to fine tune treatments by focussing on the individuals' strengths and weaknesses based on his/hers CHC profile (e.g. Proctor 2012; Niileksela & Reynolds 2014). Examples on how application of the CHC model can benefit diagnosis in and treatment of individuals with limited cognitive abilities are described in Uzieblo *et al.* (2012) and Fiorello & Primerano (2005). Furthermore, new or adjustments to intelligence tests are increasingly using the CHC model as framework, as can be seen for instance in the newest version of the WAIS, the WAIS-IV.

Methodological considerations

Although information on time of administration was available, increases or decreases in IQ scores over time were not investigated in the current study. If a potential temporal association was investigated between scores, the samples sizes would have become too small to draw tangible conclusions out of the results. Larger longitudinal studies (in forensic psychiatric patients) are needed to further investigate which factors are responsible for the temporal changes found between scores. Also, it would be interesting to investigate exploratory factors other than the Flynn effect. IQ measurements were entered randomly in the data set with regard to time of administration and analysed using that order. For example, when two WAIS-III scores were available, it was possible that the WAIS-III$^{(2)}$ score was an older score than WAIS-III$^{(1)}$. Due to the random sampling of scores, no conclusion could be made in the current sample regarding increases in IQ scores over time as a result of the Flynn effect.

The practice effect refers to an increase in IQ score that results from an individual being retested on the same instrument (Kaufman 1994). Therefore, established clinical practice is to avoid administering the same intelligence test within the same year to the same individual because it will often lead to an overestimation of a person's true intelligence (Kaufman & Lichtenberger 2006; Schalock *et al.* 2010). However, in court proceedings, it is possible that an individual is being retested within a short time period by several experts. In addition, research has shown that people with lower IQ's have less 'benefit' from practice effects (Rapport *et al.* 1997). In the current study, a number of IQ tests were readministered within a year, but the order of administration of each test is random, thereby averaging out possible practice effects.

Conclusion

The current study showed that although IQ scores are correlated within persons, stability and/or interchangeability of scores is lacking, especially in individuals with a great need for a stable assessment of intelligence (i.e. individuals with intellectual disability). Differences of 10 points and more were found between IQ assessments, with the largest differences found comparing the WAIS-III with the sGIT. Therefore, although current good practices entail reporting the confidence interval together with the IQ score, further caution in interpreting IQ scores is recommended. Additionally, all neuropsychological reports should contain information regarding the norms used and report raw scores. Uniformity in the use and reporting of intelligence measurements in forensic psychiatric patients is clearly necessary. The CHC model may serve as an important framework.

Acknowledgments

The present authors thank the participating psychiatric hospitals, namely Bierbeek, Zelzate and Rekem, and Claudia Pouls for their role in data collection.

Conflict of interest

No conflict of interest has been declared.

Source of funding

This project was sponsored by Limburg Sterk Merk (LSM) and the Public Psychiatric Care Centre Rekem (OPZC Rekem).

Correspondence

Any correspondence should be directed to Petra Habets, Knowledge Centre Forensic Psychiatric Care (KeFor) OPZC Rekem, Daalbroekstraat 106, 3621 Rekem, Belgium (e-mail: Petra.habets@opzcrekem.be).

References

American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*. 4th edn (Text Revision). American Psychiatric Association, Washington, DC.

Banister P. A., Smith F. V., Heskin K. J. & Bolton N. (1973) Psychological correlates of long-term imprisonment: I Cognitive variables. *British Journal of Criminology* **13**, 312–323.

Bellisle F. (2004) Effects of diet on behaviour and cognition in children. *The British Journal of Nutrition* **92**(Suppl 2), S227–S232.

Biles D. (1968) Test performance and imprisonment. *Australian & New Zealand Journal of Criminology* **1**, 46–58.

Bolton N., Smith F. V., Heskin K. J. & Banister P. A. (1976) Psychological correlates of long-term imprisonment: IV A longitudinal analysis. *British Journal of Criminology* **16**, 38–47.

Colom R., Lluis-Font J. M. & Andres-Pueyo A. (2005) The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: supporting evidence for the nutrition hypothesis. *Intelligence* **33**, 83–91.

Deary I. J., Whiteman M. C., Starr J. M., Whalley L. J. & Fox H. C. (2004) The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. *Journal of Personality and Social Psychology* **86**, 130–147.

Dettbarn E. (2012) Effects of long-term incarceration: a statistical comparison of two expert assessments of two experts at the beginning and the end of incarceration. *International Journal of Law and Psychiatry* **35**, 236–239.

Di Nuovo A. G., Di Nuovo S. & Buono S. (2012) Intelligent quotient estimation of mental retarded people from different psychometric instruments using artificial neural networks. *Artificial Intelligence in Medicine* **54**, 135–145.

Dutton E. & Lynn R. (2013) A negative Flynn effect in Finland, 1997–2009. *Intelligence* **41**, 817–820.

van Emmerik J. L. (2001) *At Disposal to be Treated on Behalf of the State: The Numbers [De Terbeschikkingstelling in Maat en Getal]*. Ministerie van Veiligheid en Justitie, DJI, Den Haag.

Evans I. M. (1991) Testing and diagnosis: a review and evaluation. In: *Critical Issues in the Lives of People With Severe Disabilities* (eds L. H. Meyer, C. A. Peck & L. Brown), pp. 25–44. Brookes Publishing Co, Baltimore, MD.

Fabian J. M., Thompson W. W. & Lazarus J. B. (2011) Life, death, and IQ: it's much more than just a score: understanding and utilizing forensic psychological and neuropsychological evaluations in Atkins intellectual disability/mental retardation cases. *Cleveland State Law Review* **59**, 402–430.

Fiorello C. A. & Primerano D. (2005) Research into practice: Cattell-Horn-Carroll cognitive assessment in practice: eligibility and program development issues. *Psychology in the Schools* **42**, 525–536.

Floyd R. G., Clark M. H. & Shadish W. R. (2008) The exchangeability of IQs: implications for professional psychology. *Professional Psychology: Research and Practice* **39**, 414–423.

Flynn J. R. (1984) The mean IQ of Americans: massive gains 1932 to 1978. *Psychological Bulletin* **101**, 171–191.

Flynn J. R. (2006) Tethering the elephant: capital cases: IQ, and the Flynn effect. *Psychology, Public Policy, and Law* **12**, 170–189.

Goethals J. (1981) Imprisonment and cognitive impairment. *Journal of Clinical Psychology* **37**, 418–422.

Gow A. J., Johnson W., Pattie A., Brett C. E., Roberts B., Starr J. M. & Deary I. J. (2011) Stability and change in intelligence from age 11 to ages 70, 79, and 87: the Lothian Birth Cohorts of 1921 and 1936. *Psychology and Aging* **26**, 232–240.

Heaton R. K., Gladsjo J. A., Palmer B. W., Kuck J., Marcotte T. D. & Jeste D. V. (2001) Stability and course of neuropsychological deficits in schizophrenia. *Archives of General Psychiatry* **58**, 24–32.

Jensen A. R. (1980) *Bias in Mental Testing*. Methuen, London.

Kaufman A. S. (1994) Practice effects. In: *Encyclopedia of Intelligence* (ed. Sternberg R. J.), pp. 828–833. Macmillan, New York.

Kaufman A. S. & Kaufman N. L. (1993) *Kaufman Adolescent and Adult Intelligence Test (KAIT) Manual*. American Guidance Service Inc, Circle Pines, MN.

Kaufman A. S. & Lichtenberger E. O. (2006) *Assessing Adolescent and Adult Intelligence*. Allyn & Bacon, Needham Heights, MA.

Kooreman A. & Luteijn F. (1987) *The Short Format of the Groningen Intelligence Test [Groninger Intelligentie Test: Schriftelijke Verkorte Vorm]*. Swets & Zeitlinger, Lisse.

Koyama K. I., Asakawa A., Nakahara T., Amitani H., Amitani M., Saito M., Taruno Y., Zoshiki T., Cheng K. C., Yasuhara D. & Inui A. (2012) Intelligence quotient and cognitive functions in severe restricting-type anorexia nervosa before and after weight gain. *Nutrition* **28**, 1132–1136.

Luteijn F. & Barelds D. P. H. (2004) *Git2. Groninger Intelligence Test2 [Git2. Groninger Intelligentie Test 2]*. Harcourt Assessment, Amsterdam.

Magez W., Moenaert H. & Degezelle A. (2006) *RAVEN Standard Progressive Matrices (SPM) Norms for Aduls in Flanders in Relation With the Dutch Adaption of the WAIS-III [RAVEN Standaard Progressive Matrices (SPM) Normen Voor Volwassen in Vlaanderen in Relatie met de WAIS-III Nederlandstalig Bewerking]*. CAPvzw, Brasschaat.

McBrien J. (2003) The intellectually disabled offender: methodological problems in identification. *Journal of Applied Research in Intellectual Disabilities* **16**, 95–105.

McGrew K. S. (2009) Editorial: CHC theory and the human cognitive abilities project: standing on the shoulders of the giants of psychometric intelligence research. *Intelligence* **37**, 1–10.

Moenaert H. (2006) *WAIS-III NL and Raven SPM in Flemisch Persons With Occupational Disability [WAIS-III NL en Raven SPM bij Vlaamse Arbeidsgehandicapten]*. CBWV, Leuven.

Niileksela C. R. & Reynolds M. R. (2014) Global, broad, or specific cognitive differences? Using a MIMIC model to examine differences in CHC abilities in children with learning disabilities. *Journal of Learning Disabilities* **47**, 224–236.

Proctor B. (2012) Relationships between cattell-horn-carroll (CHC) cognitive abilities and math achievement within a

sample of college students with learning disabilities. *Journal of Learning Disabilities* **45**, 278–287.

Rapport L. J., Brines D. B., Axelrod B. N. & Theisen M. E. (1997) Full scale IQ as mediator of practice effects: the rich get richer. *The Clinical Neuropsychologist* **11**, 375–380.

Raven J. C., Court J. H. & Raven J. (1983) *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Lewis, London.

Raven J., Raven J. C. & Court J. H. (1998) *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Harcourt Assessment, San Antonio, TX.

Rodgers J. L. (1998) A critique of the Flynn Effect: massive IQ Gains, methodological artifacts, or both? *Intelligence* **26**, 337–356.

Roid G. H. (2003) *Stanford-Binet Intelligence Scales*. Riverside Publishing, Itasca, IL.

Schalock R. L., Borthwick-Duffy S. A., Bradley V. J., Buntinx W. H. E., Coulter D. L., Craig E. M., Gomez S. C., Lachapelle Y., Luckasson R., Reeve A., Shogren K. A., Snell M. A., Spreat S., Tassé M. J., Thompson J. R., Verdugo-Alonso M. A., Wehmeyer M. L. & Yeager M. H. (2010) *Intellectual Disability: Definition, Classification, and Systems of Supports*. American Association on Intellectual and Developmental Disabilities, Washington, DC.

Schroth M. L. (1983) A study of aging, intelligence and problem solving. *Psychological Reports* **53**, 1271–1279.

Silverman W., Miezejeski C., Ryan R., Zigman W., Krinsky-McHale S. & Urv T. (2010) Stanford-Binet & WAIS IQ differences and their implications for adults with intellectual disability (aka mental retardation). *Intelligence* **38**, 242–248.

Smithers L. G., Golley R. K., Mittinty M. N., Brazionis L., Northstone K., Emmett P. & Lynch J. W. (2012) Dietary patterns at 6, 15 and 24 months of age are associated with IQ at 8 years of age. *European Journal of Epidemiology* **27**, 525–535.

StataCorp (2011) *Stata Statistical Software: Release 12*. StataCorp LP, College Station, TX.

Teasdale T. W. & Owen D. R. (1989) Continuing secular increases in intellgence and a stable prevalence of high intelligence levels. *Intelligence* **13**, 255–262.

Teasdale T. W. & Owen D. R. (2005) A long term rise and recent decline in intelligence test performance: the Flynn Effect in reverse. *Personality and Individual Differences* **39**, 837–843.

Teasdale T. W. & Owen D. R. (2008) Secular declines in cognitive test scores: a reversal of the Flynn Effect. *Intelligence* **36**, 121–126.

Tellegen P. (2002) The quality of the WAIS-III norms [De kwaliteit van de normen van de WAIS-III]. *De Psycholoog* **37**, 463–465.

Tulkin S. R. & Newbrough J. R. (1968) Social class, race, and sex differnences on the Raven. *Journal of Consulting and Clinical Psychology* **32**, 400–406.

Uzieblo K., Winter J., Vanderfaeillie J., Rossi G. & Magez W. (2012) Intelligent diagnosing of intellectual disabilities in offenders: food for thought. *Behavioral Sciences and the Law* **30**, 28–48.

Van Toorn B. & Bon C. (2011) The unreliability of IQ-tests. *De Psycholoog* **2011**, 44–49.

Verlinden S., Maes B. & Goethals J. (2009) *Persons With Intellectual Disability who are Forensic Psychiatric Patients [Personen met een Verstandelijke Handicap Onderhevig aan een Interneringsmaatregel]*. Steunpunt Welzijn, Volksgezondheid en Gezin, Leuven.

Wechsler D. (1955) *Manual for the Wechsler Adult Intelligence Scale WAIS*. Psychological Corporation, New York.

Wechsler D. (1970) *WAIS Dutch Adaptation. Technical Manual [WAIS Nederlandstalige Bewerking.Technische Handleiding]*. Swets & Zeitlinger, Amsterdam.

Wechsler D. (1997) *Wechsler Adult Intelligence Scale, Third ed. Administration and Scoring Manual*. Psychological Corporation, San Antonio, TX.

Wechsler D. (2000) *WAIS-III Dutch Adaptation. Technical Manual [WAIS-III Nederlandstalige Bewerking.Technische Handleiding]*. Swets & Zeitlinger, Lisse.

Wechsler D. (2005) *WAIS-III Dutch Adaptation. Technical Manual [WAIS-III Nederlandstalige Bewerking.Technische Handleiding]*. Swets & Zeitlinger, Lisse.

Whitaker S. (2008a) Intellectual disability: a concept in need of revision? *The British Journal of Developmental Disabilities* **54**, 3–9.

Whitaker S. (2008b) The stability of IQ in people with low intellectual ability: an analysis of the literature. *Intellectual and Developmental Disabilities* **46**, 120–128.

Whitaker S. (2010) Error in the estimation of intellectual ability in th elow range using WISC-IV and WAIS-III. *Personality and Individual Differences* **48**, 517–521.

Whitaker S. (2013) *Intellectual Disability: An Inability to Cope With an Intellectually Demanding World*. Palgrave MacMillan, London.

Williams R. L. (2013) Overview of the Flynn effect. *Intelligence* **41**, 753–764.

van Winkel R., Myin-Germeys I., Delespaul P., Peuskens J., De Hert M. & van Os J. (2006) Premorbid IQ as a predictor for the course of IQ in first onset patients with schizophrenia: a 10-year follow-up study. *Schizophrenia Research* **88**, 47–54.

Zampera E. (1999) Intellectual performance of chronic schizophrenic patients. *Collegium Antropologicum* **23**, 597–602.