



## CLINICAL RESEARCH ARTICLE

# Pitfalls of using IQ short forms in neurodevelopmental disorders: a study in patients with congenital heart disease

Melanie Ehrler<sup>1,2</sup>, Beatrice Latal<sup>1,2</sup>, Susanne Polentarutti<sup>1</sup>, Michael von Rhein<sup>1,2,3</sup>, Leonhard Held<sup>4</sup> and Flavia M. Wehrle<sup>1,2</sup>

**BACKGROUND:** Short forms of IQ (S-IQ) assessments are time efficient and highly predictive of the full IQ (F-IQ) in healthy individuals. To investigate the validity of S-IQs for patients with neurodevelopmental impairments, this study tested a well-established S-IQ version in patients with congenital heart disease (CHD).

**METHODS:** The Wechsler Intelligence Scale for Children, Fourth Edition was applied in 107 children with complex CHD aged 9–11 years. F-IQ and a well-established S-IQ version were calculated for each patient. The agreement between S-IQ and F-IQ was investigated across the whole spectrum of IQ scores. Finally, we tested a method to adjust IQs to resolve potential bias and validated this method in an independent sample of 55 CHD patients.

**RESULTS:** S-IQ and F-IQ correlated strongly. Nevertheless, the size of the bias correlated with the true IQ, indicating larger error at the tails of the distribution. Estimating a corrected IQ by adjusting the S-IQ with correction parameters substantially improved agreement.

**CONCLUSION:** We here report that substantial bias may underestimate low IQ scores and overestimate high ones. This bias should be considered when at-risk populations are assessed with S-IQs. Importantly, the bias can be minimized by using a correction formula.

*Pediatric Research* (2020) 87:917–923; <https://doi.org/10.1038/s41390-019-0667-2>

## INTRODUCTION

Many neurodevelopmental disorders are characterized by deficits in general intellectual abilities, reflected in lower intelligence quotients (IQs) than those of typically developing peers. For example, the mean IQ of preterm cohorts has been reported to lie about ten IQ points below that of typically developing children (see Allotey et al.<sup>1</sup> for a meta-analysis). Children with congenital heart disease (CHD) after cardiopulmonary bypass (CPB) surgery have comparable IQ deficits.<sup>2–4</sup> Low IQ scores, particularly those below internationally recognized cut-offs defining learning and intellectual disability, are associated with a higher need for therapeutic interventions and educational action.<sup>5,6</sup> Consequently, IQ assessments in children at risk for neurodevelopmental disorders can guide long-term care and ensure the timely initiation of support.<sup>7,8</sup>

In research settings, IQ assessment often complements the assessment of other neurodevelopmental domains, such as executive functions.<sup>9</sup> A common practice in such settings is to administer only a subset of the tests used for full-scale IQ (F-IQ) assessment to avoid lengthy protocols, because these may reduce performance due to fatigue or loss of motivation in patients.<sup>10,11</sup> Regression equation models have been developed to estimate IQ from various subsets of tests.<sup>11,12</sup> These short IQ (S-IQ) assessments have been proven to accurately predict the F-IQ in large samples of typically developing children and adolescents<sup>10,11</sup> and in patients with neurological disorders.<sup>13</sup> Moreover, most children with intellectual disabilities were identified correctly when S-IQ assessments were applied.<sup>14</sup> Importantly, the administration of S-IQ assessments has been

shown to reduce testing time by >50% compared to the administration of the full test.<sup>15</sup>

So far, research on the appropriateness of S-IQ assessments in neurodevelopmental disorders has been limited. Several studies have tested various S-IQ versions in patients at risk for neurodevelopmental deficits, and their findings can guide the optimal selection of subtests for specific clinical populations.<sup>16–18</sup> However, these studies have not examined whether S-IQ versions accurately measure IQ across the whole spectrum of IQ scores. In fact, there is evidence for a large bias in S-IQ scores outside the normal IQ range in healthy adults.<sup>19</sup> This bias may be particularly pronounced in neurodevelopmental disorders due to the distribution shift toward lower IQ scores.<sup>7</sup>

Consequently, this study aimed to comprehensively investigate the appropriateness of S-IQ assessment in a group of patients at risk for neurodevelopmental disorders, CHD patients. Specifically, we tested whether S-IQ assessments adequately estimate IQ across the whole spectrum of scores when IQ was assessed with the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV). Further, we present a statistical approach to minimizing potential bias between S-IQ and F-IQ and validated this in an independent cohort of CHD patients.

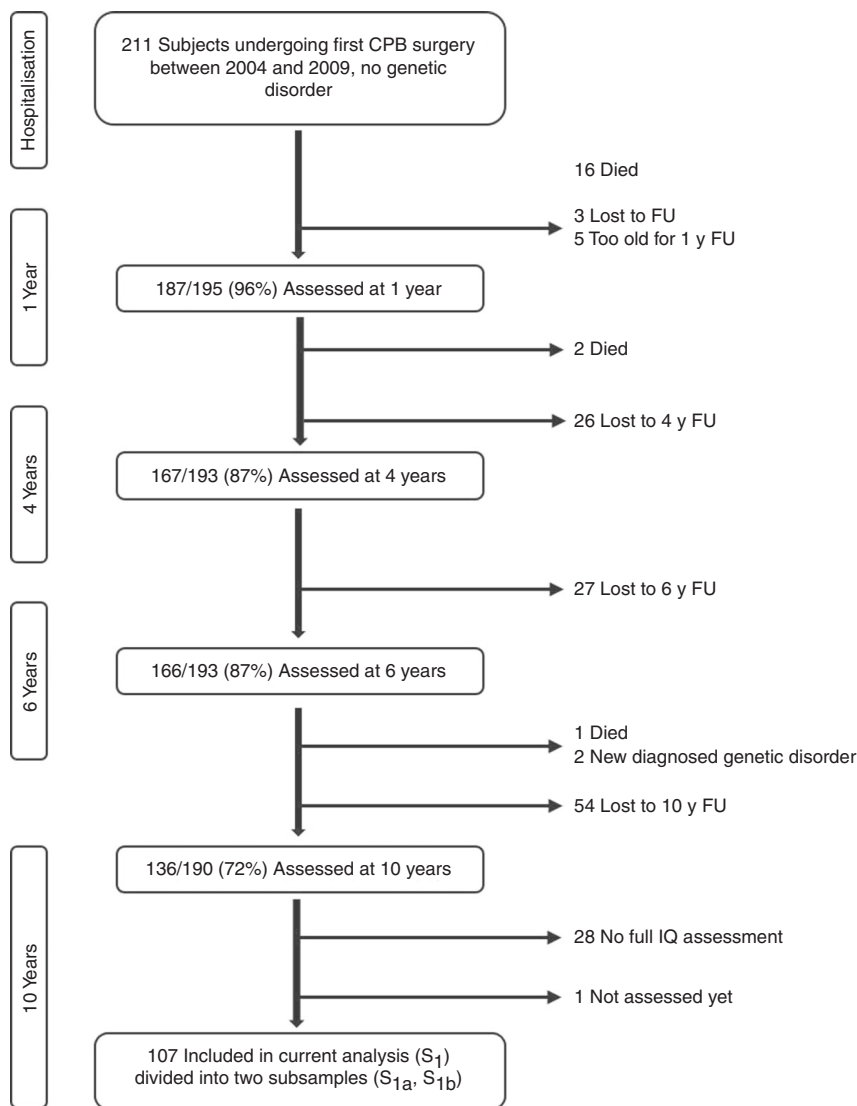
## METHODS

### Participants

Patients were recruited from the Research and Child Health Outcome cohort, which consists of 211 children who underwent CPB surgery at the University Children's Hospital Zurich,

<sup>1</sup>Child Development Center, University Children's Hospital Zurich, Zurich, Switzerland; <sup>2</sup>Children's Research Center, University Children's Hospital Zurich, Zurich, Switzerland; <sup>3</sup>Developmental Pediatrics, SPZ, Cantonal Hospital Winterthur, Zurich, Switzerland and <sup>4</sup>Department of Biostatistics, University of Zurich, Zurich, Switzerland  
Correspondence: Flavia M. Wehrle (flavia.wehrle@kispi.uzh.ch)

Received: 7 June 2019 Revised: 17 September 2019 Accepted: 16 October 2019  
Published online: 11 November 2019



**Fig. 1** Flow diagram of participant selection for sample  $S_1$ , y FU year follow-up.

Switzerland between 2004 and 2009. This prospective cohort study included 5 assessments: Before surgery and at 1, 4, 6, and 10 years of age. The current study uses the data from the 10-year neurodevelopmental assessment (see Fig. 1). Patients with a genetic or dysmorphic syndrome, CPB before study enrolment, or age >6 years at first surgery were excluded from the 10-year assessment. Of the 190 eligible patients, 136 agreed to participate. Twenty-eight patients did not complete all subtests of the WISC-IV required to calculate the S-IQ used in this study and were therefore excluded. In total, data of 107 patients was analyzed for the current study. To enable correction parameters to be developed, the total sample ( $S_1$ ) was divided into two subsamples ( $S_{1a}$ ,  $S_{1b}$ ) by random sampling.

The reproducibility of our findings was confirmed in an independent sample of patients with CHD ( $S_2$ ): We used the IQ data of 55 patients with CHD from a retrospective cohort study on CHD conducted at the University Children’s Hospital Zurich. These participants underwent CPB between 1995 and 1998 and were aged 11–16 years at the time of the study (for further details, see Schaefer et al.<sup>4</sup>).

The Ethical Committee of the Canton of Zurich, Switzerland approved both cohort studies. Written informed consent was obtained prior to the study participation from parents and from participants aged >11 years.

#### Outcome measures

IQ was assessed by well-trained professionals using the full-length German version of the WISC-IV.<sup>20</sup> Information on the heart defect was collected from medical records. Socio-economic status (SES) was estimated by rating maternal education and paternal occupation on a 6-point scale ranging from 2 to 12.<sup>21</sup>

F-IQ was calculated according to the instructions provided in the manual of the WISC-IV and adjusted for age. The S-IQ was estimated according to a regression equation method described by Waldmann.<sup>11</sup> We selected a subtest combination that retains the conceptual structure of the WISC-IV to preserve construct validity. This includes one subtest from each of the four indices: working memory, processing speed, reasoning, and verbal comprehension. One-per-index-subtest combinations are often used in research to estimate IQ.<sup>22–24</sup> Of these subtest combinations, we chose the equation with the highest reliability according to Waldmann<sup>11</sup> ( $R^2 = 0.911$ ; subtests: letter–number sequencing, symbol search, matrices, vocabulary).

#### Statistical analyses

Sample characteristics—sex, age, SES, and type of heart defect (cyanotic, acyanotic)—are reported for both the full sample ( $S_1$ ), the two subsamples ( $S_{1a}$ ,  $S_{1b}$ ) and the validation sample ( $S_2$ ). Two-sided Welch’s  $t$  test allowing for unequal variance and a two-

**Table 1.** Sample characteristics of the full sample ( $S_1$ ), its subsamples ( $S_{1a}$ ,  $S_{1b}$ ), and the validation sample ( $S_2$ ).

Characteristics	Full sample ( $S_1$ )	Subsample ( $S_{1a}$ )	Subsample ( $S_{1b}$ )	Validation sample ( $S_2$ )	$P$ value ( $S_1$ , $S_2$ )	$P$ value ( $S_{1a}$ , $S_{1b}$ )
Age (mean, range)	10.2 (9–11)	10.2 (9–11)	10.3 (9–11)	13.7 (11–16)	<0.001	0.8
Sex (% female)	35%	32%	37%	56%	0.01	0.7
SES (mean, range)	8.5 (4–12)	8.5 (4–12)	8.5 (4–12)	7.8 (3–12)	0.06	0.9
Cyanotic CHD (%)	36%	31%	40%	47%	0.2	0.5

sampled chi-squared test were used to investigate the comparability of the two subsamples ( $S_{1a}$  and  $S_{1b}$ ), the full ( $S_1$ ) and validation samples ( $S_2$ ).

#### Correlation and agreement between S-IQ and F-IQ

The correlation between the F-IQ and the S-IQ was examined using Spearman's correlation. We conducted two types of analyses to investigate the agreement between the F-IQ score and the S-IQ. First, we examined the overall agreement: We tested for a significant mean difference between the S-IQ and the F-IQ (i.e., the mean bias) by using a paired two-sided  $t$  test. Further, the proportion of participants with a bias  $\leq 5$  IQ points (within 95% confidence interval (CI)) was calculated. Second, we analyzed whether the bias between the two IQ measurements (F-IQ, S-IQ) depends on the IQ itself (i.e., whether the bias is equal across the whole IQ spectrum). Thus we correlated the bias with the estimated true IQ using Spearman's correlation.<sup>25</sup> The bias was defined as the absolute difference between the F-IQ and the S-IQ, and the true IQ was estimated from the mean of the S-IQ and F-IQ because using either value separately would cause a statistical artifact.<sup>26</sup>

Agreement was further tested by dichotomizing the data above a clinically relevant cut-off into normal ( $IQ \geq 85$ ) or below into abnormal IQ ( $IQ < 85$ ). In order to test for agreement, a two-sampled chi-squared test was conducted using IQ type (F-IQ, S-IQ) as predictor and IQ score (normal, abnormal) as outcome. Further, the proportion of patients who fell into the same category for the S-IQ and the F-IQ was calculated.

#### Correction method for S-IQ

To minimize the bias between the F-IQ and the S-IQ, we applied a statistical method to estimate a corrected IQ. Therefore, the full sample ( $S_1$ ) was split into two subsamples by means of random sampling ( $S_{1b}$ ,  $S_{1a}$ ). We calculated a linear regression model for the subsample  $S_{1a}$ , using S-IQ as predictor and F-IQ as outcome ( $F-IQ = \beta^0 + \beta^1 \times S-IQ$ ). The intercept ( $\beta^0$ ) and slope ( $\beta^1$ ) were then used in the correction model to predict the corrected IQ of the subsample  $S_{1b}$ :

$$\text{Corrected IQ}_{1b} = \beta^0_{1a} + \beta^1_{1a} * S-IQ_{1b}$$

Two measures for quantifying the agreement with the F-IQ were calculated: the root mean square error (RMSE) and the mean absolute error (MAE). These were calculated for both the S-IQ and the corrected IQ and were then compared. A lower RMSE and MAE in the corrected IQ than in the S-IQ was interpreted as improvement of agreement.

#### Validation of the correction method

Finally, we validated the correction method in another independent sample of CHD patients. We estimated a corrected IQ in the validation sample ( $S_2$ ) by implementing our correction parameters obtained from the subsample  $S_{1a}$  ( $\beta^0$ ,  $\beta^1$ ) to the S-IQ of the validation sample ( $S_2$ ).

All the statistical analyses were performed using the R computing environment, version 3.5.3.<sup>27,28</sup> All analyses were performed with an  $\alpha$  level of 0.05.

## RESULTS

### Sample characteristics

Table 1 provides sample characteristics of the full sample ( $S_1$ ), the two subsamples ( $S_{1a}$ ,  $S_{1b}$ ), and the validation sample ( $S_2$ ). Evidence for differences between the full and validation samples were found for age ( $t(55.7) = 16.60$ ,  $P < 0.001$ , 95% CI: from 3.03 to 3.86), and sex ( $\chi^2(1) = 6.21$ ,  $P = 0.01$ ). There was no difference for the type of heart defect (cyanotic and acyanotic;  $\chi^2(1) = 1.64$ ,  $P = 0.2$ ) and only weak evidence for a difference in SES ( $t(92.9) = 1.88$ ,  $P = 0.06$ , 95% CI: from  $-0.04$  to 1.47). The two subsamples ( $S_{1a}$ ,  $S_{1b}$ ) did not differ in patient characteristics (see Table 1).

Table 2 provides F-IQ, S-IQ, and the mean difference between the F-IQ and the S-IQ (i.e., bias) for the full sample, the two subsamples, and the validation sample. There was strong evidence of a difference in the mean F-IQ between the full sample and the validation sample ( $S_1$ ,  $S_2$ :  $t(90.6) = 3.09$ ,  $P = 0.003$ , 95% CI: from 2.88 to 13.23). There was no evidence of a difference in mean F-IQ between the two subsamples ( $S_{1a}$ ,  $S_{1b}$ :  $t(104.0) = 0.22$ ,  $P = 0.8$ , 95% CI: from  $-4.60$  to 5.76).

### Correlation and agreement between S-IQ and F-IQ

The analysis revealed a very strong correlation between the S-IQ and the F-IQ in the full sample ( $S_1$ :  $r = 0.95$ ,  $P < 0.001$ , Fig. 2a). A two-sided paired  $t$  test provided no evidence of a difference between S-IQ and F-IQ ( $S_1$ :  $t(106) = 1.50$ ,  $P = 0.1$ , 95% CI: from  $-0.28$  to 2.04). The mean difference between S-IQ and F-IQ (i.e., mean bias) in the full sample was 0.9 IQ point (Table 2). Furthermore, 81% of all participants had a bias  $\leq 5$  IQ points.

When considering IQ scores across the whole IQ spectrum, our findings revealed a very strong correlation between bias and the true IQ score ( $S_1$ :  $r = -0.62$ ,  $P < 0.001$ ). A Bland-Altman plot of difference (Fig. 2b) illustrates the underestimation of S-IQ for true IQ scores below the mean and the overestimation of S-IQ for true IQ scores above the mean.

Categorizing S-IQ and F-IQ scores as normal ( $IQ \geq 85$ ) or abnormal ( $IQ < 85$ ) revealed no evidence for a difference in the distribution between the categories in the full sample ( $S_1$ :  $\chi^2(1) = 0.65$ ,  $P = 0.4$ ). The proportion of patients of the full sample ( $S_1$ ) who fell into the same category for the S-IQ and the F-IQ was 93%.

### Correction method for S-IQ

The correction of the S-IQ of subsample  $S_{1b}$  with the parameters ( $\beta^0$ ,  $\beta^1$ ) from the subsample  $S_{1a}$  resulted in a mean corrected IQ of 95.7 (SD = 12.2, Table 2). The correlation between the F-IQ and the corrected IQ remained strong ( $r = 0.94$ ,  $P < 0.001$ , Fig. 3b; for comparison: correlation between F-IQ and uncorrected S-IQ: Fig. 3a). Both RMSE (RMSE<sub>S-IQ</sub> = 6.0; RMSE<sub>corrected IQ</sub> = 4.1) and MAE (MAE<sub>S-IQ</sub> = 4.5; MAE<sub>corrected IQ</sub> = 3.3) were lower in the corrected IQ than in the S-IQ. This is evidence for successful correction of the bias using the correction formula provided in this study.

To test the stability of the correction method, correction parameters ( $\beta^0$ ,  $\beta^1$ ) were also extracted from subsample  $S_{1b}$  and applied to subsample  $S_{1a}$ . The results were identical (see Supplementary Fig. S1 for details). Correction parameters for all one-per-index-subtest combinations with an  $r > 0.9^{11}$  and for

**Table 2.** Comparison of F-IQ and S-IQ for all samples and comparison of F-IQ and corrected IQ for the subsample ( $S_{1b}$ ) and the validation sample ( $S_2$ ).

	Mean	Standard deviation	Range	Mean difference <sup>a</sup>
Full sample ( $S_1$ )				
F-IQ	95.7	13.4	53–120	
S-IQ	94.8	17.2	49.2–126.0	−0.9
Subsample ( $S_{1a}$ )				
F-IQ	95.3	14.0	53–120	
S-IQ	94.6	18.1	49.2–123.2	−0.7
Subsample ( $S_{1b}$ )				
F-IQ	95.9	12.9	61–118	
S-IQ	95.0	16.4	53.3–126.0	−0.9
	95.7	12.2	64.8–118.6	−0.2
Corrected IQ				
Validation sample ( $S_2$ )				
F-IQ	103.7	16.7	44–131	
S-IQ	107.5	21.5	25–142	−3.8
	104.9	15.9	43.8–130.6	1.2
Corrected IQ				

<sup>a</sup>Mean bias between S-IQ and F-IQ

further combinations consisting of 2–6 subtests are listed in Supplementary Table S2a and Fig. S2b.

#### Validation of the correction method

The correction method was applied in a validation sample ( $S_2$ ) to further confirm its stability. Adjusting the S-IQ of the validation sample ( $S_2$ ) with the correction parameters ( $\beta^0$ ,  $\beta^1$ ) from the subsample ( $S_{1a}$ ) resulted in a mean corrected IQ of 104.9 (SD = 15.9, Table 2). The correlation between the F-IQ and the corrected IQ remained strong ( $r = 0.93$ ,  $P < 0.001$ , Supplementary Fig. S3). Both RMSE (RMSE<sub>S-IQ</sub> = 7.8; RMSE<sub>corrected IQ</sub> = 4.6) and MAE (MAE<sub>S-IQ</sub> = 6.3; MAE<sub>corrected IQ</sub> = 3.4) were lower in the corrected IQ than in the S-IQ. This is evidence for successful correction of the bias in the validation sample.

## DISCUSSION

In this study, we investigated whether short forms of IQ assessments (S-IQ) accurately estimate IQ scores in patients at risk for neurodevelopmental disorders. Our results indicate that, while the correlation between the F-IQ assessment and S-IQ is high overall, IQ scores at the tails of the distribution are not accurately measured when using S-IQ assessments. Using our correction method, we were able to considerably reduce this bias.

In this analysis of children and adolescents with CHD, a population with frequent neurodevelopmental impairments, we calculated F-IQ scores from the full WISC-IV and S-IQ scores from a well-established four-subtest version.<sup>11</sup> First, we demonstrated a strong correlation between the S-IQ and the F-IQ scores. These findings are in line with previous studies on both healthy participants<sup>10,11</sup> and patients with neurological disorders.<sup>13,17</sup> On examining the agreement between these two IQ measures, we observed a substantial variation in bias across the spectrum of IQ: S-IQ scores were underestimated in the lower spectrum of IQ and overestimated in the higher spectrum. This bias was replicated in various other subtest combinations (see Supplementary Table S2a and Fig. S2b) and in an independent sample of CHD patients. A previous study in healthy adults that used other estimates of the

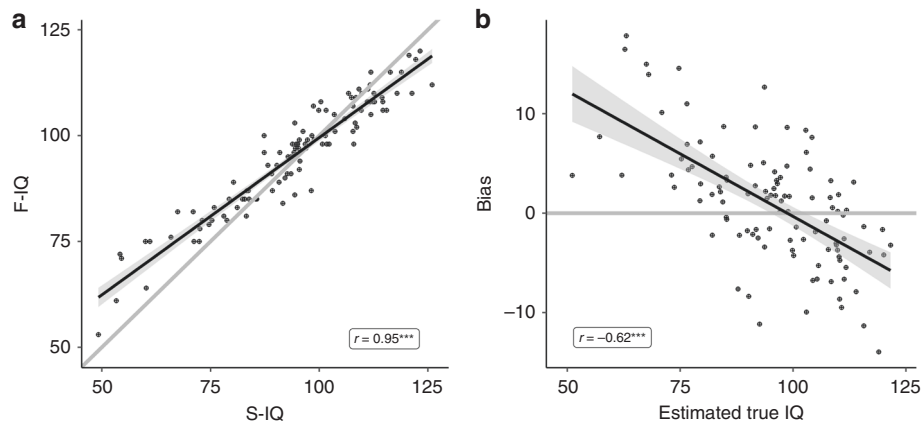
Wechsler Adult Intelligence Scale, Third Edition also reported larger bias at the tails of the distribution.<sup>19</sup>

Time-efficient assessments of IQ by means of S-IQ methods are particularly valuable in research settings because IQ is often not the primary outcome of interest but serves merely as a covariate or patient characteristic to allow inferences to be drawn about potential selection bias and generalizability from the assessed cohort.<sup>29</sup> Using short forms allows more time for the detailed assessment of developmental domains of interest, such as executive function<sup>4,9</sup> and memory performance.<sup>30</sup> However, the usefulness of S-IQs has long been a matter of debate.<sup>31,32</sup> In particular, concerns have been raised whether the construct validity assessed in a sample with a full-length assessment holds true when only a subsample of tests is applied.<sup>11,32</sup> For this reason, we aimed to retain the best construct validity possible by using a one-subtest-per-index short form. This approach has frequently been used in previous studies (e.g.,<sup>22–24</sup>). Furthermore, it has been debated whether S-IQ scores are valid in samples whose IQ distributions differ from those of the general population, for example, in patients with population-specific cognitive deficits. Indeed, several studies have reported considerable mean IQ differences between S-IQ and F-IQ versions in cohorts of patients with epilepsy,<sup>17</sup> traumatic brain injury,<sup>16</sup> neurologic disorders,<sup>13</sup> and borderline intellectual disabilities.<sup>18</sup> However, these studies only investigated overall mean differences between the S-IQ and the F-IQ, not the changing size and direction of bias across the whole spectrum of IQ scores. Our finding that bias is largely restricted to the tails of the distribution suggests that the significant mean differences observed in those studies likely occurred owing to large biases in low IQ scores.

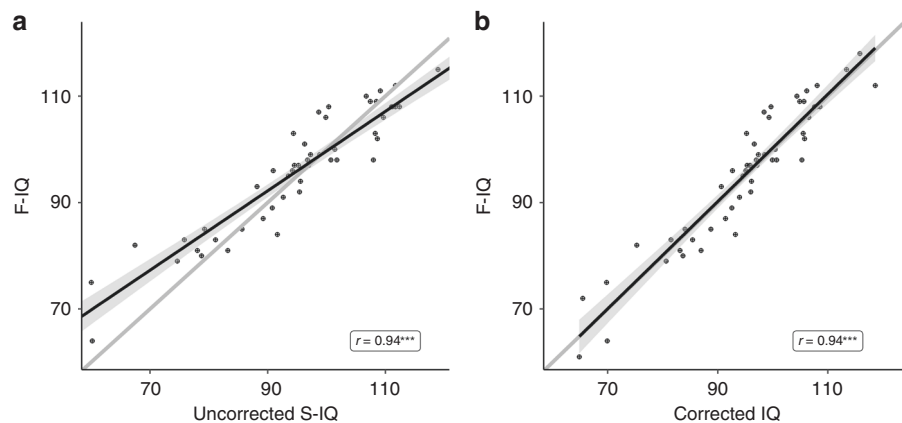
Importantly, the bias observed between the mean S-IQs and F-IQs of a population only provides very limited information on the agreement between the two measures because it does not reflect the spectrum of scores. The concept of agreement has been discussed extensively in the past. Unfortunately, there is confusion about appropriate concepts to test the agreement between two measurements, and this can lead to misinterpretation of results.<sup>33</sup> Appropriate methods to investigate agreement consider the degree of bias across the whole spectrum of scores and therefore provide substantial information about the quality of a measurement.<sup>25</sup> Therefore, this study examined bias across a whole spectrum of IQ scores by correlating the bias with the estimated true IQ, as suggested by Altman and Bland.<sup>25</sup>

The fact that the bias occurs predominantly at the tails of the distribution may reflect a statistical phenomenon, called regression to the mean, which causes one measurement to be closer to the population mean than another measurement of the same entity. This is especially apparent for values outside the normal range.<sup>34</sup> Regression to the mean has previously been discussed as a problem affecting the interpretation of IQ scores.<sup>35–37</sup> In particular, F-IQ scores outside the normal range tend to be closer to the mean than the corresponding S-IQ scores, since sporadic extreme subtest scores are weighted less in the F-IQ. Our findings show that it is particularly important to consider the distribution of the data when S-IQ scores of populations at risk for neurodevelopmental impairments are used in statistical analyses.

To overcome the bias between F-IQ and S-IQ measures, we tested a correction method to resolve this problem. Correction parameters were obtained from one subsample and were applied to the other subsample to estimate a corrected IQ. We observed an improved agreement between the two measurements: The RMSE of the S-IQ was as high as six before the correction, indicating low agreement.<sup>11</sup> The RMSE of the corrected IQ was substantially reduced (RMSE = 4.1), reflecting a considerable improvement in agreement. These findings underline that correcting S-IQ scores may well improve the validity of an estimate, especially when its distribution differs from the normative sample.



**Fig. 2** **a** Correlation of S-IQ and F-IQ. **b** Correlation of bias and estimated true IQ, both in the full sample ( $S_1$ ).<sup>25</sup> Dots = participants. Black line = linear regression line. Gray line = perfect agreement. Gray shadow = 95% CI.  $r$  = Spearman's correlation. \*\*\* $P < 0.001$ .



**Fig. 3** **a** Correlation between F-IQ and S-IQ and **b** F-IQ and corrected IQ in the subsample  $S_{1b}$ . Dots = participants. Black line = linear regression line. Gray line = perfect agreement. Gray shadow = 95% CI.  $r$  = Spearman's correlation. \*\*\* $P < 0.001$ .

The correction formula proved useful both when subdividing the full sample into two random subsamples and when correcting the S-IQ of an independent validation sample of CHD patients ( $S_2$ ). Indeed, very similar improvements of IQ estimations were found in the validation sample that was different in regard to age, SES, and sex. The mean F-IQ of the validation sample and its distribution differed strongly from the original full sample ( $S_1$ ). In fact, the mean IQ of the validation sample was above average, and its distribution was left-skewed, in all likelihood due to selection bias.<sup>4</sup> Nevertheless, the correction formula also worked in this very different sample of patients with CHD. This implies that the correction formula we present here may also reduce bias in other samples with abnormal distributions of IQ, irrespective of the direction of the skewness. This may be of particular interest for the correction of bias in typically developing control groups, as sampling bias is a well-known problem with high-functioning individuals.<sup>38</sup>

If IQ estimates serve to describe the proportion of individuals with clinically relevant IQ deficits, S-IQ scores may be categorized into normal ( $IQ \geq 85$ ) and abnormal IQ ( $IQ < 85$ ). Our findings reveal that a high proportion of participants were categorized correctly with the S-IQ. These findings are in line with a study by Murray and colleagues,<sup>14</sup> who used a cut-off score of 70 and successfully identified children with intellectual disabilities using an S-IQ estimation. This method is easy to apply when concerns about bias are present. However, categorizing data results in a loss of valuable information.<sup>39</sup> Therefore, it is important to consider why the IQ has been measured and decide whether using categories only is appropriate.

Taken together, using S-IQs and the presented correction formula may be useful in research settings to provide an appropriate estimation of general intellectual abilities, especially when IQ serves as a secondary outcome. In clinical settings, the full-length assessment of IQ is, however, preferable as only this provides a detailed assessment of the global IQ and the respective indices. This is required to identify deficits in specific cognitive domains and to initiate appropriate interventions.

#### LIMITATIONS

Our study has several limitations worth mentioning. We investigated a cohort of patients with CHD as one example of populations at risk for neurodevelopmental impairments. Therefore, our findings cannot directly be transferred to other at-risk populations. Other studies should investigate clinical populations with a higher proportion of individuals with very low IQ (e.g., severe intellectual disabilities) in order to confirm the validity of the correction formula. Also, future studies should validate our findings in populations with different profiles of cognitive impairment. Nevertheless, the cognitive profile of children born very preterm is akin to that of patients with CHD,<sup>40</sup> so we propose that the correction formula presented here could also be applied to correct the S-IQ scores of preterm-born children.

We did not have a control sample assessed with the full WISC-IV. Therefore, we could not compare the correlation and agreement of F-IQ with S-IQ between typically developing children and adolescents and those with CHD. However, a study by Spinks and colleagues<sup>19</sup> in healthy adults has also reported a higher

proportion of large bias (>5 IQ points) at the tails of the distribution than within the normal range. This finding indicates that the bias in the outer spectrum may also occur in cohorts of healthy individuals. The proportion of individuals with IQ scores <85 is rather low in typically developing cohorts, though, so this may have less bearing than in at-risk populations.

This study only investigated estimates of IQ by means of equation models but did not consider other methods (e.g., Wechsler Abbreviated Scale for Intelligence, Second Edition, Wechsler 2011). This was mainly because the equation models published by Waldmann<sup>11</sup> are frequently used in research on German-speaking populations with various clinical characteristics (e.g.,<sup>22–24</sup>). Further, we only investigated a variety of four-subtest versions designed to minimize testing time while retaining the conceptual structure of the WISC-IV. Thus we cannot draw any conclusion regarding the agreement and potential correction methods in S-IQ obtained from other estimates.

Our analyses on S-IQ were done with data of patients who conducted the full-length WISC-IV assessment. Applying a short-form version requires considerably less time,<sup>15</sup> which may lead to less fatigue and better motivation,<sup>13</sup> resulting in a better outcome. This has to be considered when short forms and the presented correction formula are used. However, the short-form estimates by Waldmann<sup>11</sup> were also developed from data of the full-length assessment. Our analyses are, therefore, in line with the original work of Waldmann.

## CONCLUSION

S-IQs measured in a sample of children with CHD with a well-established four-subtest version revealed a lack of agreement with F-IQs at the tails of the distribution, namely, in low and high performers of the WISC-IV. This warrants particular consideration when investigating populations whose IQ scores differ from those of the general population. Bias between S-IQ and F-IQ scores can be diminished either by dichotomizing the data into scores above or below 85 or by using a correction formula presented in this study.

## ACKNOWLEDGEMENTS

We thank the children and parents for their participation. We thank Professor Dr. med. Walter Knirsch for his help regarding cardiac variables. This work was supported by the Swiss Heart Foundation and the Else Kröner-Fresenius Foundation. The sponsors had no influence on study design, the collection, analysis, and interpretation of data, the writing of the report, or the decision to submit the paper for publication.

## AUTHOR CONTRIBUTIONS

Each author has met the *Pediatric Research* authorship requirements. M.E. carried out the analyses and interpretation of the data and drafted and revised the manuscript. B.L. and M.v.R. conceived and designed the study and critically revised the manuscript for important intellectual content. S.P. collected the data and critically revised the manuscript for important intellectual content. L.H. contributed substantial statistical support in order to adequately analyze the data and critically revised the manuscript for important intellectual content. F.M.W. made substantial contributions to the concept and design of the analyses and to the interpretation of the results. She further critically reviewed and revised the manuscript for important intellectual content. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

## ADDITIONAL INFORMATION

The online version of this article (<https://doi.org/10.1038/s41390-019-0667-2>) contains supplementary material, which is available to authorized users.

**Competing interests:** The authors declare no competing interests.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## REFERENCES

1. Allotey, J. et al. Cognitive, motor, behavioural and academic performances of children born preterm: a meta-analysis and systematic review involving 64 061 children. *BJOG Int. J. Obstet. Gynaecol.* **125**, 16–25 (2018).
2. Bellinger, D. C. et al. Adolescents with tetralogy of Fallot: neuropsychological assessment and structural brain imaging. *Cardiol. Young* **25**, 338–347 (2015).
3. Naef, N. et al. Neurodevelopmental profiles of children with congenital heart disease at school age. *J. Pediatr.* **188**, 75–81 (2017).
4. Schaefer, C. et al. Neurodevelopmental outcome, psychological adjustment, and quality of life in adolescents with congenital heart disease. *Dev. Med. Child Neurol.* **55**, 1143–1149 (2013).
5. O'Brien, G. Defining learning disability: what place does intelligence testing have now? *Dev. Med. Child Neurol.* **43**, 570–573 (2001).
6. Wieland, J. & Zitman, F. G. It is time to bring borderline intellectual functioning back into the main fold of classification systems. *BJPsych Bull.* **40**, 204–206 (2016).
7. Latal, B. Neurodevelopmental outcomes of the child with congenital heart disease. *Clin. Perinatol.* **43**, 173–185 (2016).
8. Marino, B. S. et al. Neurodevelopmental outcomes in children with congenital heart disease: Evaluation and management. *Circulation* **126**, 1143–1172 (2012).
9. Calderon, J. et al. Executive function and theory of mind in school-aged children after neonatal corrective cardiac surgery for transposition of the great arteries. *Dev. Med. Child Neurol.* **52**, 1139–1144 (2010).
10. Crawford, J. R., Anderson, V., Rankin, P. M. & MacDonald, J. An index-based short-form of the WISC-IV with accompanying analysis of the reliability and abnormality of differences. *Br. J. Clin. Psychol.* **49**, 235–258 (2010).
11. Waldmann, H. C. Kurzformen des HAWIK-IV: statistische bewertung in verschiedenen anwendungsszenarien. *Diagnostica* **54**, 202–210 (2008).
12. Crawford, J. R., Allan, K. M. & Jack, A. M. Short-forms of the UK WAIS—R: Regression equations and their predictive validity in a general population sample. *Br. J. Clin. Psychol.* **31**, 191–202 (1992).
13. van Ool, J. S. et al. Accuracy of WISC-III and WAIS-IV short forms in patients with neurological disorders. *Dev. Neurorehabil.* **21**, 101–107 (2018).
14. Murray, A. L., McKenzie, K. & Murray, G. C. An evaluation of the performance of the WISC-IV eight-subtest short form with children who may have an intellectual disability. *J. Intellect. Dev. Disabil.* **41**, 50–53 (2016).
15. Ryan, J. J., Glass, L. A. & Brown, C. N. Administration time estimates for Wechsler Intelligence Scale for Children-IV Subtests, Composites, and Short Forms. *J. Clin. Psychol.* **63**, 309–318 (2007).
16. Donders, J., Elzinga, B., Kuipers, D., Helder, E. & Crawford, J. R. Development of an eight-subtest short form of the WISC-IV and evaluation of its clinical utility in children with traumatic brain injury. *Child Neuropsychol.* **19**, 662–670 (2013).
17. Hrabok, M., Brooks, B. L., Fay-McClumont, T. B. & Sherman, E. M. S. Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV) short-form validity: a comparison study in pediatric epilepsy. *Child Neuropsychol.* **20**, 49–59 (2014).
18. Van Duijvenbode, N., Didden, R., Van Den Hazel, T. & Engels, R. C. M. E. Psychometric qualities of a tetrad WAIS-III short form for use in individuals with mild to borderline intellectual disability. *Dev. Neurorehabil.* **19**, 26–30 (2016).
19. Spinks, R. et al. IQ estimate smackdown: comparing IQ proxy measures to the WAIS-III. *J. Int. Neuropsychol. Soc.* **15**, 590 (2009).
20. Peterman, F. & Peterman, U. Hawik-IV. *Kindh. Entwickl.* **17**, 71–75 (2008).
21. Largo, R. H., Pfister, D., Molinari, L., Kundu, S. & Lipp, A. D. G. Significance of prenatal, peri-natal and postnatal factors in the development of AGA preterm infants at five to seven years. *Dev. Med. Child Neurol.* **31**, 440–456 (1989).
22. Groch, S. et al. Memory cueing during sleep modifies the interpretation of ambiguous scenes in adolescents and adults. *Dev. Cogn. Neurosci.* **17**, 10–18 (2016).
23. Liechti, M. D. et al. First clinical trial of tomographic neurofeedback in attention-deficit/hyperactivity disorder: evaluation of voluntary cortical control. *Clin. Neurophysiol.* **123**, 1989–2005 (2012).
24. Wehrle, F. M. et al. Very preterm adolescents show impaired performance with increasing demands in executive function tasks. *Early Hum. Dev.* **92**, 37–43 (2016).
25. Altman, D. G. & Bland, J. M. Measurement in medicine: the analysis of method comparison studies. *Statistician* **32**, 307–317 (1983).
26. Gill, S., Beevers, D. G., Zezulka, V. & Davies, P. Relation between initial blood pressure and its fall with treatment. *Lancet* **1**, 567–569 (1985).
27. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2018)
28. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer Verlag New York, 2016).

29. Steiner, P. M., Cook, T. D., Shadish, W. R. & Clark, M. H. The importance of covariate selection in controlling for selection bias in observational studies. *Psychol. Methods* **15**, 250–267 (2010).
30. Cassidy, A. R., Newburger, J. W. & Bellinger, D. C. Learning and memory in adolescents with critical biventricular congenital heart disease. *J. Int. Neuropsychol. Soc.* **23**, 627–639 (2017).
31. Donders, J. Using a Short Form of the WISC-III: sinful or smart? *Child Neuropsychol.* **7**, 99–103 (2001).
32. Kaufman, J. C. & Kaufman, A. S. Time for the changing of the guard: a farewell to short forms of intelligence tests. *J. Psychoeduc. Assess.* **19**, 245–267 (2001).
33. Kottner, J. & Streiner, D. L. The difference between reliability and agreement. *J. Clin. Epidemiol.* **64**, 701–702 (2011).
34. Bland, M. J. & Altman, D. G. Regression towards the mean. *Stat. Notes* **30**, 1499 (1994).
35. Basso, M. R., Bornstein, R. A., Roper, B. L. & McCoy, V. L. Limited accuracy of premorbid intelligence estimators: a demonstration of regression to the mean. *Clin. Neuropsychol.* **14**, 325–340 (2000).
36. Larrabee, G. J. FORUM association between IQ and neuropsychological test performance: commentary on Tremont, Hoffman, Scott and Adams (1998). *Clin. Neuropsychol.* **14**, 139–145 (2000).
37. Veiel, H. O. F. & Koopman, R. F. The bias in regression-based indices of premorbid IQ. *Psychol. Assess.* **13**, 356–368 (2001).
38. Russell, E. W. Norming subjects for the Halstead Reitan battery. *Arch. Clin. Neuropsychol.* **20**, 479–484 (2005).
39. Royston, P., Altman, D. G. & Sauerbrei, W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat. Med.* **25**, 127–141 (2006).
40. Easson, K. et al. A comparison of developmental outcomes of adolescent neonatal intensive care unit survivors born with a congenital heart defect or born preterm. *J. Pediatr.* **207**, 34–41 (2018).