



UvA-DARE (Digital Academic Repository)

The validity of WISC-V profiles of strengths and weaknesses

de Jong, P.F.

DOI

[10.1177/07342829221150868](https://doi.org/10.1177/07342829221150868)

Publication date

2023

Document Version

Final published version

Published in

Journal of Psychoeducational Assessment

License

CC BY

[Link to publication](#)

Citation for published version (APA):

de Jong, P. F. (2023). The validity of WISC-V profiles of strengths and weaknesses. *Journal of Psychoeducational Assessment*, 41(4), 363-379.
<https://doi.org/10.1177/07342829221150868>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

The Validity of WISC-V Profiles of Strengths and Weaknesses

Journal of Psychoeducational Assessment
2023, Vol. 41(4) 363–379
© The Author(s) 2023



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/07342829221150868
journals.sagepub.com/home/jpa



Peter F. de Jong¹ 

Abstract

The Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V; [Wechsler, 2014](#)) provides a general intelligence score, representing *g*, and five index scores, reflecting underlying broad factors. Within person differences between the overall performance across subtests and index scores, denoted as index difference scores, are often used to examine profiles of strengths and weaknesses. In this study, the validity of such profiles was examined for the Dutch WISC-V. In line with previous studies, broad factors explained little variance in index scores. A simulation study showed that variation in index difference scores also reflected little broad factor variance. The simulation study further revealed that, as a consequence, a significant discrepancy between an index score and overall performance was accompanied in only 40%–74% of the cases by a discrepancy on the underlying broad factor. Overall, these results provide little support for the validity and thereby clinical use of WISC-V profiles.

Keywords

WISC-V, strengths-based assessment, validity, exceptionalities/disabilities

Intelligence is probably the most widely assessed ability around the world ([Evers et al., 2012](#)). In children it is used to inform a range of educational decisions including eligibility for special services, placement in secondary education, underperformance in the classroom and the diagnosis of several developmental disorders that are relevant in the context of school ([Benson et al., 2020](#)).

Intelligence scores or IQ scores are usually determined with a battery of (sub)tests. Additionally, variation in performance across subtests or composites of subtest [i.e. profiles of strengths and weaknesses] are often considered ([Kranzler et al., 2020](#)). However, the clinical use of such profiles has been hotly debated for many years. Some have argued that further scrutiny of a profile, like a ‘detective’ ([Kaufman, 1994](#)), can provide additional information that is relevant to understanding an individual’s intellectual functioning and needs for treatment. Others have questioned the reliability, stability and validity of such profiles ([Macmann & Barnett, 1997](#); [McGill](#)

¹Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands

Corresponding Author:

Peter F. de Jong, Research Institute of Child Development and Education, University of Amsterdam, PO box 15780, 1001 NG Amsterdam, The Netherlands.

Email: P.F.deJong@uva.nl

et al., 2018; Watkins & Canivez, 2004), and tried to debunk the ‘myth of the master detective’ (Macmann & Barnett, 1997) or advised to ‘just say no to subtest analysis’ (McDermott et al., 1990). The release of the fifth edition of the Wechsler Intelligence Scale for Children (WISC-V) seems to have given a new impetus to this debate (Canivez et al., 2017; Dombrowski et al., 2018; McGill et al., 2018). In this study, I consider the validity of profiles of strengths and weaknesses on the basis of the Dutch version of the WISC-V (Wechsler, 2018).

The WISC-V is based on the Cattell-Horn-Carroll model (CHC-model) of cognitive abilities, which McGrew (2009) claims to be the most encompassing taxonomy of cognitive abilities. The CHC-model resulted from merging the Cattell-Horn model with Carroll’s Three Stratum model, two models of the structure of cognitive abilities which are often regarded as similar although only the latter model includes *g* (Canivez & Youngstrom, 2019). The CHC model is a hierarchical model with a general factor at the highest level affecting approximately 16 broad abilities at a second level, which in turn each affect a range of specific cognitive abilities at the lowest level.

The WISC-V consists of 16 subtests that are assumed to measure *g*, and five broad abilities of the CHC model: Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory and Processing Speed. The major scores are derived from only 10 subtests; a Full Scale IQ (FSIQ) based on seven subtests and the General Index Score (GIS-index) based on 10 subtests. In addition, five index scores can be derived, each based on two of the subtests. The index scores are presumed to reflect the five broad abilities. Furthermore, index difference scores can be computed by subtracting the GIS-index (or FSIQ) from an index score. A significant discrepancy between an index score and the GIS-index (or FSIQ) indicates a strength in case the index score is higher than the GIS-index, or a weakness, in case the index score is lower (Wechsler, 2014; 2018).

Structure of the WISC-V

In a number of studies Canivez and colleagues have addressed several problems with the interpretation of the five index scores (Canivez et al., 2017; Canivez, et al., 2018; Dombrowski et al., 2015; 2018; 2022; Watkins & Canivez, 2022). One problem concerns the factor structure of the WISC-V. The appropriateness of the five factor model is important as it underlies the index scores which form the basis for the interpretation of strengths and weaknesses in cognitive abilities. The five factor model reported in the technical manual of the American version has redundancies as shown by re-analyses of data (e.g., Dombrowski et al., 2015; Dombrowski et al., 2019) as well as by a simulation study (Dombrowski, et al., 2021). For example, the second order *g* factor and the broad factor Fluid Reasoning are identical and the model implied correlation of Fluid Reasoning with Visual Spatial is .88. In a series of studies Canivez and colleagues showed that a model with four broad factors provides a better and more parsimonious description of the 16 subtests as well as the 10 primary subtests of the US version of the WISC-V (Canivez, et al., 2016; 2017; Dombrowski, et al., 2018; Dombrowski et al., 2019). In their four factor solutions, the Visual-Spatial and Fluid Reasoning factors merged together. Similar results were reported for the French and the German version of the WISC-V (Lecerf & Canivez, 2018; Pauls & Daseking, 2021).

Another problem with the validity of the index scores, mentioned by Canivez and colleagues, concerns the extent to which these scores reflect the underlying broad factor. This problem is considered in the present study for the scores of the Dutch version of the WISC-V. In addition, also the validity of the index difference scores is examined as these scores form the basis of profiles of strengths and weaknesses.

Validity of Index Scores

Canivez and colleagues have repeatedly highlighted the limited amount of variance in the index scores that can be attributed to the associated underlying broad factor, denoted here as the lack of validity of the index scores (e.g., [Canivez et al., 2017](#); [McGill et al., 2018](#); [Watkins & Canivez, 2022](#)). Most of the variance of an index score is accounted for by the *g* factor and not by the broad factor that it is presumed to reflect. A review by [McGill et al. \(2018\)](#) shows that the broad factors on average account for less than 25% of the variance, with the processing speed index at one extreme (18%–52%) and fluid reasoning at the other (0%–6.6%). Therefore, Canivez and colleagues have concluded that the index scores of the WISC-V are of questionable value for clinical decisions ([Canivez et al., 2017](#); [Watkins & Canivez, 2022](#)).

Most studies on the (un)interpretability of WISC-V index scores have followed the logic of factor analysis and decomposed the variance of the index scores into three independent sources: the *g* factor, the broad factor and unique variance. But for a particular subtest, the unique variance consists of measurement error and specific variance. Measurement error indicates the unreliability of the subtest. The specific variance of a test reflects one or more additional abilities that are required for a particular subtest but are unrelated to the specific abilities needed for the other subtest associated with the index ([Brunner et al., 2012](#); [Schneider, 2013](#)). The specific abilities of both subtests are also a source of variance for the index score. Usually, these specific abilities are assumed to reflect very specific, and mostly uninteresting, features of the task. In some cases, however, there might be a substantive interpretation of the subtest-specific variance. For example, the working memory index is determined by two span tasks, Digit Span and Picture Span. These task probably (partly) represent different aspects of the working memory system ([Baddeley, 2012](#)), with Digit Span depending more on the verbal subsystem and Picture Span more on the visual-spatial subsystem. Thus, the amount of task specific variance in the index scores can be of interest, but is still largely unknown.

Validity of Index Difference Scores

As said, a strength or weakness on an index does not concern the index score as such, but a difference between an index score and the GIS-index (or FSIQ). Accordingly, an important question in the use of profiles is the extent to which index difference scores actually reflect the underlying broad factor. At first sight these difference index scores might fare better than the index scores because these difference scores are corrected for the influence of the *g* factor. However, two caveats should be mentioned. First, the absence of the *g* factor in a difference index score comes at a cost, as difference scores can be unreliable and will certainly be less reliable than the index scores ([Cronbach & Furby, 1970](#); [Watkins & Canivez, 2022](#)). Second, there is a relation between the sources of variance in the index and the index difference score. That is, the amount of broad factor variance in an index score affects the amount of broad factor variance in an index difference score. However, how a lack of broad factor variance in an index score affects the amount of variance accounted for by the broad factor in the index difference scores of the WISC-V is not yet known. It is also unclear to what extent these difference index scores reflect other sources of variance, such as subtest specific factors.

Even if an index difference score reflects an insufficient amount of variance of the underlying broad factor, the meaning of such an outcome remains somewhat elusive. Knowledge of the amount of such variance might not clarify the relation between an observed discrepancy (strength or weakness) and the likelihood of a discrepancy of similar magnitude on an underlying factor. Put differently, for practitioners it seems more apt to ask how often a significant weakness or strength

on an index score can be attributed to a weakness or strength on the broad factor and/or to the specific factors that are believed to underlie a particular index score.

Current Study

The main aim of the current study was to examine the validity of the index and index difference scores of the Dutch WISC-V. With respect to the five index scores, the amounts of variance accounted for by the various sources, that is the g factor, the broad factor and the specific factors, were derived from the five factor solution as reported in the technical manual (Wechsler, 2018). To put these amounts of variance into perspective, they were compared to the variance described by these sources in the US version of the WISC-V (Wechsler, 2014). Following previous studies, it was expected that most of the variance in the index scores is accounted for by the g factor, with smaller contributions of the broad factor and the specific factors. Next, the validity of the five index difference scores was assessed through a simulation study. The amount of variance captured by the broad and specific factors was expected to be somewhat higher in the difference scores than in the index scores, as the index difference scores do not reflect the g factor. The data of the simulation study were also used to examine how often a weakness or strength on an index score, that is a significant positive or negative index difference score, was accompanied by a similar weakness or strength on the underlying broad factor.

Method

Variance Decomposition of Index Scores

Computation of the contributions of variance by the various sources underlying the index scores was based on Figure 1 for the Dutch WISC-V and on Dombrowski et al. (2018, Figure 1, p. 92) for the US WISC-V. The reliabilities of the primary subtests came from the Dutch and US technical manuals (Wechsler, 2014, 2018).

Variance decomposition of the index scores comprised two steps. First, the variance of each subtest was decomposed into variance accounted for by the g factor, variance accounted for by the broad factor, error variance and specific variance. The proportion of variance accounted for by the g factor was computed by multiplying the factor loading of a subtest on its broad factor by the factor loading of the broad factor on the g factor and then taking the square root of this product (Brunner et al., 2012). The proportion of variance contributed by the broad factor after the g factor variance was accounted for, was computed by subtracting the g factor variance from the square of the factor loading of the subtest on the broad factor. The proportion of error variance due to measurement error was obtained by subtracting the reliability of the subtest as reported in the technical manual from one. Finally, the proportion of specific variance was computed by subtracting the proportion of error variance from the residual subtest variance, which was computed by subtracting the squared loading of the test on the broad factor from one.

In the second step the total variance of an index score was decomposed. Coefficient Omega hierarchical scale was used to compute the unique proportion of variance that each of the factors (g factor, broad factor, specific factors and errors) accounted for in the variance of the index score (Brunner et al., 2012; Reise, 2012). Although, Omega hierarchical scale is usually considered a reliability estimate of a scale, it essentially expresses the amount of variance described by one source of variance divided by the total amount of variance described by all sources of variance of a score, in this case an index score. The computations were based on an adaptation of Formula 3 in Brunner et al. (2012, p. 825). The adaptation was to split the residual variance e_i into (test) specific variance s_i and measurement error e_i (see Appendix A). For valid clinical interpretation of an index

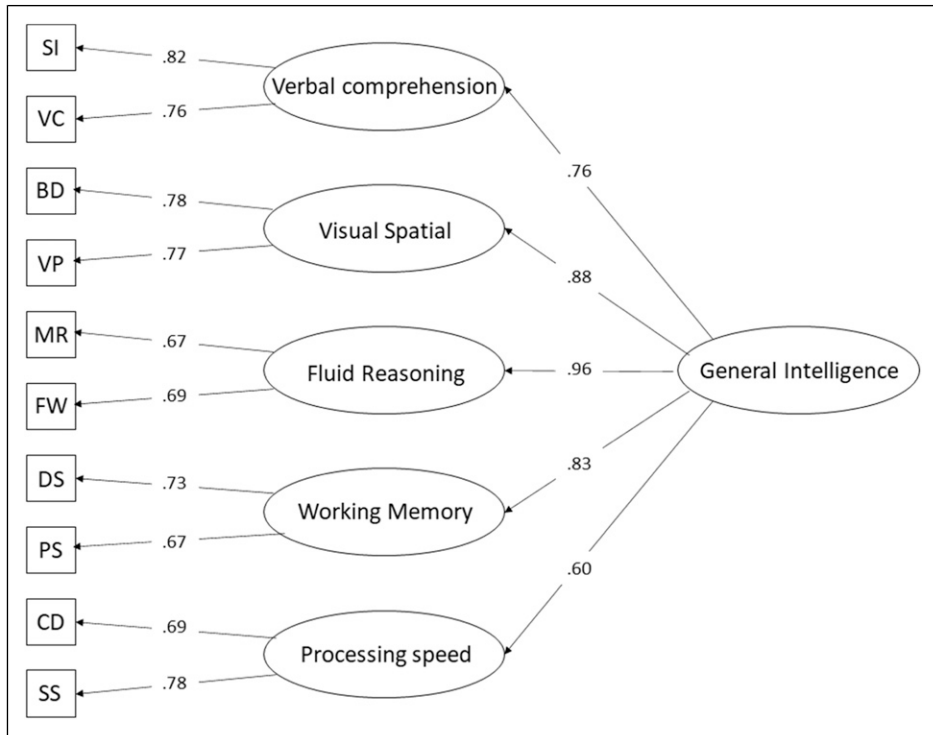


Figure 1. Hierarchical factors model of the Dutch WISC-V with standardized coefficients (information taken from Figure 6.2 in Wechsler [2018]). SI = Similarities, VC = Vocabulary, BD = Block Design, VP = Visual Puzzles, MR = Matrix Reasoning, FW = Figure Weights, DS = Digit Span, PS = Picture Span, CD = Coding, SS = Symbol Search.

score, Omega hierarchical scale for the broad factor should be at a minimum of .50, but .75 is to be preferred (Reise et al., 2013).

Variance Decomposition of Index Difference Scores

Simulation Procedure. Following the logic of the hierarchical factor model in Figure 1, each subtest can be decomposed into four independent factors: The g factor, the broad factor, the specific factor and an error factor. The full factor model of the 10 primary subtest consists of 26 independent variables: One g factor, 5 broad factors, 10 specific factors and 10 error factors. The R program (R Development Core Team, 2016) was used to generate a sample of 1,000,000 cases from a multivariate normal distribution of 26 independent variables with a mean of zero and a variance of one (Beaujean, 2018; Miciak et al., 2018). The code for the simulation is listed in Appendix B.

Generated scores were not transformed, and were not rounded to whole numbers. The z-score metric was used in all analyses because a) sources of variance in the difference scores are not dependent on metric and b) the use of the z-score metric will not affect the conclusions that follow from the results.

For each case, subtest scores were computed as the weighted sum of the underlying factors of the subtest, that is the general factor, the broad factor involved in the subtest, its specific factor and its error factor. The weights were the square root of the proportions of variance accounted for in a particular subtest (see section Validity of index scores above). For a particular subtest, these

weights are identical to the factor loadings of the subtest on its underlying factors. Then, the index scores were computed as the average of the two subtest scores and the GIS index as the average of the 10 primary subtests. Finally, five index difference scores were computed by subtracting the GIS index from each index score.

Sources of Variance in Index Difference Scores. To determine the sources of variance in the five index difference scores, each of these scores was regressed on the assumed underlying sources of variance of the index. Each difference index score was regressed on seven types of factors. Four factors were related to the particular index score: 1) the *g* factor, 2) the broad factor that the index score is assumed to reflect, 3) the subtest specific variance of the two subtests that are the basis of the index, and 4) the errors of these two subtests. Three types of factors were unrelated to the particular index, that is all other 1) broad factors (4 in total), 2) test specific factors (8 in total) and 3) subtest errors (8 in total). Together, proportions of variance of all related and unrelated underlying factors add up to one.

The factors were subsequently added to the regression model to determine the change in R^2 . This change indicated how much variance a particular type of factor added. Note that the particular order of inclusion of the factors was irrelevant, because all factors were independent, that is, had a correlation of zero, which is a consequence of the assumption underlying the factor model of the tests.

Classification Procedure of Observed and Underlying Weaknesses or Strengths

The data of the simulation study were also used to examine the relation between an observed discrepancy on an index score and discrepancies on the underlying factors of the index difference score. The classification of cases with and without a discrepancy between the index score and the GIS index was based on the index difference scores. Note that a discrepant index score means that the index difference score differs significantly from zero. Cut-off scores for discrepancies depended on the significance level of the difference between index and GIS-index. In the current study, significance levels of .05 and .01 were examined. The cut-off score that belongs to each level of significance of each particular index score was derived from the technical manual (Wechsler, 2018). The discrepancies between the GIS-index and an index score in z-scores ranged from 0.68 to 0.75 at an alpha level of .05 and from 0.81 to 0.90 at an alpha level of .01.

Next, cases with and without discrepancies on the underlying factors of the difference index scores, that is the broad factor, two test specific factors and two errors, were distinguished. Discrepancies on these five underlying factors were based on the same z-score cut-offs used to determine a discrepancy (strength or weakness) on an index score. There are 32 possible combinations of discrepancies on the five underlying factors. Three broader categories were considered: 1) The percentage of cases with an underlying discrepancy on the broad factor, 2) the percentage of cases with a discrepancy on one or both task specific factors only and 3) the percentage of cases with discrepancies on one or both errors only. The remaining possibilities concerned discrepancies on combinations of task specific and error factors which are difficult to interpret and were therefore lumped into the category 'other'. The computations were done separately for strengths and weaknesses.

Results

The results are presented in three sections. The first two sections contain the results on the sources of variance in the index and index differences scores, respectively. The last section presents the

results on the relation between observed strengths and weaknesses on index scores and the discrepancies on the factors that underlie the index difference scores.

Sources of Variance in the Index Scores

Each index score is based on two subtests. Therefore, before turning to the index scores, the sources of variance for the subtests for both the Dutch and the US version are reported in [Table 1](#). Percentages of variance related to the *g* factor were somewhat higher for the Dutch version for the subtests related to the factors Verbal Comprehension, Visual-Spatial and Fluid Reasoning and were somewhat lower for the subtests related to Working Memory and Processing Speed. The opposite pattern was found for the variance accounted for by the broad factors. However, the general pattern of results for the Dutch and US versions was the same. With the exception of the subtests of Processing Speed, the *g* factor accounted for about [three](#) times more variance in the subtest scores than the broad factor (Dutch version: 40% vs. 15%; US version: 44% vs. 12%). The subtest-specific factors accounted for substantial amounts of variance across subtest. On average, subtest specific factors captured 31% of the test variance in the Dutch version and 30% of the test variance in the US version. On several subtests, the specific factor explained more score variance than the associated broad factor.

The sources of variance for the index scores are presented in [Table 2](#). The proportion of systematic variance, the total variance minus the error variance, ranged from 88% to 93%. Following the recommendation of [Watkins and Canivez \(2022\)](#), suggesting a minimum of 80%, the reliability of these measures can be qualified as sufficient. Omega hierarchical scale was used to compute variance accounted for by the broad factors. Evidently, Omega hierarchical scale values for the broad factors of the US version were identical to those reported by [Canivez et al. \(2017\)](#). However, in the current study, a further distinction was made between test specific

Table 1. Proportion of Variance per Source for each of the 10 Subtests of the Dutch and US WISC-V.

Factor/Subtest	Sources of variance							
	Dutch version				US version			
	G	B	S	E	G	B	S	E
Verbal comprehension								
Similarities	.388	.284	.158	.170	.474	.196	.198	.130
Vocabulary	.334	.244	.302	.120	.486	.204	.181	.130
Visual spatial								
Block design	.471	.137	.192	.200	.468	.112	.262	.160
Visual puzzles	.459	.134	.257	.150	.493	.117	.312	.110
Fluid reasoning								
Matrix reasoning	.414	.035	.391	.160	.453	.007	.408	.130
Figure eights	.439	.037	.434	.090	.453	.017	.478	.060
Working memory								
Digit span	.367	.166	.317	.150	.419	.191	.302	.090
Picture span	.309	.140	.391	.160	.291	.129	.428	.150
Processing speed								
Coding	.171	.305	.384	.140	.127	.363	.330	.180
Symbol search	.219	.389	.252	.140	.179	.511	.121	.190

Note. G = *g*-factor; B = Broad factor; S = Test specific factor; E = Error.

Table 2. Proportion of Variance per Source per Index and General Score for the Dutch and US WISC-V.

Score	Sources of variance							
	Dutch version				US version			
	G	B	S	E	G	B	S	E
Index								
Verbal comprehension	.444	.325	.142	.089	.572	.239	.113	.077
Visual spatial	.581	.169	.140	.109	.593	.139	.182	.086
Fluid reasoning	.583	.050	.282	.086	.620	.013	.303	.065
Working memory	.453	.205	.238	.104	.467	.211	.242	.080
Processing speed	.253	.450	.206	.091	.192	.548	.143	.117
General								
GIS	.808	.086	.071	.034	.822	.081	.067	.030
FSIQ	.780	.079	.096	.045	.812	.062	.090	.037

Note. 1. GIS = General Index score; FSIQ = Full Scale IQ; G = G-factor; B = Broad factor; S = Test specific factors; E = Error.
 Note. 2. For the index scores proportions of variance reflected by factors G, B, S and E concern the variance of the factors related to the particular index. For the GIS index and FSIQ proportions concern the sum of proportions for all factors of a certain type (G, B, S and E) that are involved in the 10 subtests.

variance and error variance. As for the subtests, the pattern of results for the index scores of the Dutch and US versions was highly similar. Variance due to the broad factors was low (mostly below 30%) and, with the exception of the Processing Speed index in the US version, far below the minimal level of 50%, needed to warrant clinical interpretation of an index score (Reise et al., 2013). The broad factor Fluid Reasoning hardly explained any variance in the index score. For all index scores, the amount of variance explained by the broad factor was substantially lower than the amount accounted for by the *g* factor. Importantly, for Fluid Reasoning and Working Memory the amount of variance captured by the broad factors was even lower than that accounted for by subtest specific factors. In contrast, the GIS-index and the FSIQ seemed, as intended, to be a good reflection of the *g* factor, which accounted for about 80% of the variance. This is above the preferred level of 75% for clinical interpretation (Reise et al., 2013).

Sources of Variance in the Index Difference Scores

The sources of variance of the index difference scores were derived from the simulated data. First, however, the quality of these data was examined. To this end, in the simulated data, each of the index scores was regressed on its underlying sources of variance. Regression analyses were also conducted for the FSIQ and the GIS index. The outcomes of these regression analyses were compared to the Omega hierarchical scale values reported for the Dutch version of the WISC-V in Table 2. The results were virtually identical. A very small difference, ranging from .001 to .003, was found for 7 out of the 28 proportions.

Next, a regression analysis per index difference score was conducted to determine the proportion of variance captured by the various underlying factors. The results are presented in Table 3.

As expected, the *g* factor accounted for negligible amounts of variance. The amount of variance captured by the broad factors was higher than in the index scores but still below the minimum of 50% (Reise et al., 2013). Similar to the index scores, substantial amounts of variance in the index difference scores could be attributed to the subtest specific factors. In the difference index score for

Table 3. Proportion of Variance per Source per Difference Index score for the Dutch WISC-V.

Difference index	Sources of variance							
	Related to index score					Unrelated to index score		
	G	B	S	E	B + S	Bu	Su	Eu
Verbal comprehension	.000	.477	.208	.131	.685	.076	.074	.034
Visual spatial	.028	.297	.246	.191	.525	.109	.090	.039
Fluid reasoning	.014	.085	.490	.148	.575	.133	.084	.046
Working memory	.000	.297	.350	.151	.629	.095	.072	.036
Processing speed	.050	.491	.226	.099	.717	.052	.054	.027

Note. G = G-factor; B = Broad factor; S = Test specific factors; E = Error; B + S = Sum of Broad and Test Specific factors; Bu = Broad factors not related to difference index; Su = Specific factors not related to difference index; Eu = errors of the subtests not related to difference index.

Fluid Reasoning, even 49% of the variance was due to the specific factors of the two subtests. In the fifth column of [Table 3](#), the variance captured by the sum of the variances of the broad factor and the two subtest specific factors (B + S) is reported. Using this sum assumes that both the broad factor and the test specific factors are relevant for the domain of cognitive abilities reflected by an index difference score. Even taking this lenient approach, however, these joint sources of variance only capture between 54% and 72% of the variance. Although this exceeds the minimum amount of 50%, it is still below the level of 75% preferred for clinical interpretation ([Reise et al., 2013](#)).

Relations Between Observed and Underlying Strengths and Weaknesses

Groups with and without a discrepancy on a particular index score were distinguished for strengths and weaknesses separately. For each group, the percentage of cases was computed with an underlying discrepancy on a) the broad factor, b) one or both task specific factors only and c) one or both errors only. As expected, the results were virtually identical, because the difference index scores were normally distributed, and thus symmetrical. The difference between the percentages for strengths and weaknesses ranged from 0% to 0.3%. For ease of presentation only the exact percentages for weaknesses are presented in [Table 4](#).

Several of the results in [Table 4](#) are noteworthy. First, the results are hardly affected by the level of significance used to determine a discrepancy. Second, a discrepancy on an index score reflected a discrepancy on the underlying broad factor in 40%–74% of the cases. For Fluid Reasoning, a discrepancy on the observed index score was not accompanied by a weakness on the underlying broad factor in approximately 60% of the cases. Verbal Comprehension and Working Memory fared somewhat better. Still, for one in three cases, the discrepancy between the index and the GIS-score could not be attributed to the broad factor. Third, in 8%–20% of the cases a discrepancy was solely due to subtest specific factors. Finally, turning to cases without a discrepant index score, it appeared that approximately 30% of these cases (about one in three) did nevertheless have a discrepancy on the underlying broad factor. The results show that a discrepancy on the underlying broad factor was about 1.3 to 2.4 times more likely in cases with than in cases without a discrepant index score. The average likelihood was 1.88. According to [Chen et al. \(2010\)](#), these likelihoods mostly indicate a small effect. When taking a lenient approach by adding discrepancies on the broad and subtest specific factors (last column of [Table 4](#)), the average likelihood even decreased to 1.5.

Table 4. Percentages of cases with a weakness on underlying factors in the group of cases with and without a weakness on an index score for two levels of significance.^a

Index	Weakness	Significant	Weakness on Underlying Factors				
			B	S-only	E-only	Other	B + S-only
Verbal comprehension	Yes	.01	67.1	8.1	5.6	19.2	75.2
	Yes	.05	67.3	7.8	5.5	19.5	75.1
	No	.01	30.1	14.5	12.8	42.6	44.6
	No	.05	33.8	14.2	12.6	39.4	48.0
Visual spatial	Yes	.01	53.4	10.4	8.6	27.6	63.8
	Yes	.05	54.8	10.0	8.4	26.9	64.8
	No	.01	29.0	14.4	13.5	43.1	43.3
	No	.05	32.8	14.1	13.2	39.9	46.9
Fluid reasoning	Yes	.01	40.4	18.8	6.1	34.7	59.2
	Yes	.05	42.8	17.2	6.2	33.8	60.0
	No	.01	28.5	17.3	12.6	41.6	45.8
	No	.05	32.0	16.8	12.1	39.1	48.8
Working memory	Yes	.01	58.4	11.8	5.4	24.3	70.2
	Yes	.05	59.4	11.2	5.3	24.2	70.6
	No	.01	29.9	16.3	12.8	41.1	46.2
	No	.05	33.7	15.7	12.3	38.4	49.4
Processing speed	Yes	.01	73.3	7.2	3.4	16.0	80.5
	Yes	.05	73.3	6.9	3.3	16.4	80.2
	No	.01	30.6	16.1	12.9	40.4	46.7
	No	.05	34.1	15.5	12.3	38.1	49.6

^aReported are the exact results for a weakness. The results for a strength are virtually identical.

Note. B = weakness on B factor; S = weakness on one or both of the two subtest specific factors; E = weakness on one or both of the two subtest errors; Other = weakness on other combination of factors, mostly combinations of specific factors and errors; B + S-only = sum of B and S.

Discussion

The clinical use of profiles has been hotly debated within the domain of intelligence testing (McGill et al., 2018; Watkins & Canivez, 2022), as well as outside this domain, for example for the diagnosis of learning disabilities (Fletcher & Miciak, 2017). Many studies have been conducted that question the reliability and validity of patterns of strengths and weaknesses (Canivez et al., 2017; Macmann & Barnett, 1997; Miciak et al., 2018; Watkins et al., 2022). The current study examined the validity of a profile of strengths and weaknesses for the Dutch WISC-V. The results replicate and extend previous findings on the validity of WISC-V profiles.

As in studies on the US version of the WISC-V (e.g., McGill et al., 2018), most of the variance in the subtest and index scores of the Dutch WISC-V was accounted for by the *g* factor. The broad factors that indexes, and the associated subtest scores, are assumed to reflect, explained little variance in these scores. Especially the subtest and index scores for Fluid Reasoning, Visual Spatial and Working Memory were hardly determined by their broad factor. As an extension to previous studies, a distinction was made between subtest specific variance and error variance, for the Dutch as well as the US version of the WISC-V. The results showed that, in both the Dutch and US versions, for the majority of the subtest scores, and about two out of five index scores, task specific factors accounted for more variance than the broad factor. In all, as in earlier studies (e.g.,

McGill, et al., 2018), the subtest and index scores of the WISC-V hardly seem a reflection of the broad factors, after the *g* factor is removed.

On the basis of similar findings, Canivez and colleagues (e.g., Canivez et al., 2017; Dombrowski et al., 2018; Watkins & Canivez, 2022) have repeatedly concluded that index scores are not sufficiently valid to be used for clinical purposes. Although this conclusion might be warranted, it does not seem to follow directly from the lack of broad factor variance reflected by the index scores. First, the crucial assumption of Canivez and colleagues is that the validity of an index score should be assessed after *g* factor variance has been deleted. In contrast, it could be argued that the heavy involvement of the *g* factor in the various subtest and index scores might not be a problem for their interpretation. For example, a vocabulary test is arguably the best reflection of vocabulary knowledge, despite the fact that the test score might be substantially related to *g*. The two subtests constituting the Working Memory index clearly encompass the domain of working memory. Moreover, in some taxonomies of cognitive abilities, *g* is not even involved (Canivez & Youngstrom, 2019) or an emergent property of a dynamic system of reciprocal interactions among cognitive abilities, as in the mutualism theory (van der Maas et al., 2006; 2019). According to these theories, there would be no reason to partial out the *g* factor variance that index scores have in common. Second, as argued in the Introduction, the main issue concerns the interpretation of index difference scores as these form the basis of profiles of strengths and weaknesses. The *g* factor is hardly involved in these scores. In all, it might be debated that the substantial amount of *g* factor variance in the index scores hampers their interpretation.

Irrespective of the taxonomy or theory of cognitive abilities, the heavy involvement of the *g* factor in the index scores does become a problem when profiles of strengths and weaknesses are considered based on index differences scores. The current results show that, as in the index scores, the broad factors accounted for relatively little variance in these scores. For Fluid Reasoning, this was even less than 10%. However, unlike the index scores, the major problem with the index difference scores is that the remaining variance is hard to interpret. Depending on the particular index difference score, between 50% and 91.5% of the variance was due to a mixture of subtest specific factors and errors, as well as broad factors deemed to be irrelevant for the particular index (see Table 3). As a result, these difference scores contain insufficient systematic true score variance. Thereby, these scores are of little use for clinical interpretation.

Taking a more lenient approach, one might argue that also subtest specific factors should be taken into account to judge the validity of a difference index. In this approach, the index score is not merely conceived as the reflection of the common variance of the subtests but reflects the level of performance in a particular domain. Even taking this lenient approach, however, several difference index scores, especially the Fluid Reasoning and Visual Spatial difference indexes, hardly capture the minimally required amount of variance, and none account for the preferred amount to warrant a valid clinical interpretation (Reise et al., 2013). Moreover, though some subtest specific factors might have a substantive interpretation, such as for Digit and Picture Span, for most subtest specific factors the theoretical interpretation is largely unknown.

The consequence of a lack of validity of index difference scores was further examined by considering the extent to which a significant discrepancy on an index (strength or weakness) could be attributed to an accompanying discrepancy on the underlying broad and test specific factors. The criteria for a discrepancy were taken from the Dutch manual of the WISC-V (Wechsler, 2018), making the results directly relevant for clinicians. The results provide further insight in what it means when difference index scores are insufficiently determined by the broad factors. Especially, the Fluid Reasoning difference index hardly reflected the broad factor. As a result, in the simulated data a discrepancy on this index was accompanied by a discrepancy on the underlying broad factor in only 40% of the cases. Similarly, only 50% of the cases with a discrepancy on the Visual-Spatial index appeared to have an underlying discrepancy on this broad factor. Discrepancies on the other

indexes fared a little better. But overall, about one in two to one in four cases with a discrepancy on an index did not have an accompanying discrepancy on the underlying broad factor. These results were hardly affected by the level of significance used to determine a discrepancy. A higher significance level did result in a lower percentage of cases with a discrepancy, but did not further ensure that a discrepancy was accompanied by a discrepancy on the underlying broad factor.

Finally, a comparison was made between cases with and without a discrepancy on a particular index. Given the somewhat arbitrary cut-off to determine a discrepancy, it was to be expected that a substantial number of cases without a discrepancy on an index, did nevertheless have a discrepancy on the underlying broad factor. A weakness on an index only increased the likelihood of an underlying discrepancy on the broad factor by a factor 1.5 to 2. Such a small effect is in line with the finding that many leaning problems have multiple probabilistic causes (Pennington, 2006; van Bergen et al., 2014). However, with respect to the implications of strengths and/or weaknesses in an IQ profile for a specific learning disorder, two likelihoods are actually at stake. One is the likelihood of a learning disorder given a weakness on a certain index, and the second is the likelihood that the weakness is related to poor performance on the broad factor. When both likelihoods are rather small, it becomes difficult to tie a learning problem of a particular person to a weakness on a broad factor.

Limitations

All results of the current study are based on the Dutch factor model of the WISC-V with five first order factors (Wechsler, 2018). In this model, similar to the US model (Dombrowski et al., 2015, 2017), the first order factors Visual Spatial and Fluid Reasoning seem hard to distinguish (see Figure 1). It might be regarded a limitation that the current study was not based on the proper factor model, probably the four factor model in which the Visual Spatial and Fluid Reasoning tests form one factor. However, the five factor model currently forms the basis of the index scores used by clinicians. Moreover, as the results show, the redundancy of the five factor model is largely responsible for the little variance that can be attributed to the broad factor in the Visual Spatial and Fluid Reasoning index and difference index scores. The choice for a four factor model might not be a much better solution. The more general point is that index difference scores reflect little variance of the broad factors if the first order factors in the factor model are strongly correlated with the second order *g* factor. In a four factor solution the correlations among the factors are also (too) high (Canivez et al., 2017).

A further limitation of the use of the Dutch factor model of the WISC-V could be that the results, especially of the simulation study, might not generalize to versions of the WISC-V used in other countries. However, the factor models of the various versions tend to be highly similar (e.g., Canivez et al., 2017; Lecerf & Canivez, 2018; Pauls & Daseking, 2021). Moreover, the current study showed that the sources of variance in the subtest and index scores of the Dutch and US version are highly similar as well. As the results on the index scores directly translate into the index differences scores, it seems highly likely that a simulation study based on the US factor model of the WISC-V will give similar results and leads to the same conclusions.

The Use of Profiles of Strengths and Weaknesses

The current study, as with previous studies, strongly suggests that patterns of strengths and weaknesses lack sufficient validity for clinical use. Nevertheless, clinicians continue to use the WISC profiles. In closing, a number of reasons are considered that might be responsible for this continuing use in clinical practice. One reason is that particular profiles have often been observed in groups of individuals with specific disorders (van Iterson et al., 2015; Toffalini et al., 2017).

Possibly, clinicians might not realize that profiles observed at the group level are often not representative for substantial numbers of individuals belonging to such a group. Another reason is probably that the use of profiles is still taught to students who will eventually enter clinical practice (Farmer et al., 2021). A final reason could be that strengths and weaknesses on the WISC-V are automatically provided by the publisher to the users who are, subsequently, tempted to interpret them. As a result, the use of profiles often concerns post hoc interpretations, whereas as a proper diagnostic process should start with hypotheses to explain the problem at hand. Such hypotheses could involve a weakness or strength across the various indexes of IQ, although in many instances an absolute strength or weakness might be more important than a relative. But even in the case of an hypothesis about a discrepancy, which is important and can be posed before an IQ test battery is administered, it seems doubtful that the hypothesis can be properly tested given the lack of validity of the scores that are available to signal strengths and weaknesses.

Appendix A

Omega hierarchical scale, ω_h , denotes the proportion of variance that can be accounted for by the broad factor in the index score. According to Brunner et al. (2012, p. 825), ω_h can be defined as

$$\omega_h = \frac{\left(\sum_{i=1}^p \lambda_{ij} \right)^2}{\sum_{j=1}^k \left[\left(\sum_{i=1}^p \lambda_{ij} \right)^2 \right] + \sum_{i=1}^p e_i}$$

where λ_{ij} denotes the standardized factor loading of subtest Y_i on factor j and e_i the standardized variance of the unique factor related to subtest Y_i . J is the index for the factors involved in an index (the g -factor and the broad factor) and p indicates the particular subtests, here two. When e_i is split in task specific variance, s_i , and error, ϵ_i , ω_h becomes

$$\omega_h = \frac{\left(\sum_{i=1}^p \lambda_{ij} \right)^2}{\sum_{j=1}^k \left[\left(\sum_{i=1}^p \lambda_{ij} \right)^2 \right] + \sum_{i=1}^p s_i + \sum_{i=1}^p \epsilon_i}$$

Omega hierarchical scale for the proportion of variance accounted for by the subtest specific variance can be defined as

$$\omega_h = \frac{\sum_{i=1}^p s_i}{\sum_{j=1}^k \left[\left(\sum_{i=1}^p \lambda_{ij} \right)^2 \right] + \sum_{i=1}^p s_i + \sum_{i=1}^p \epsilon_i}$$

Omega hierarchical scale for the proportion of variance accounted for by the subtest error variance can be defined as

$$\omega_h = \frac{\sum_{i=1}^p \epsilon_i}{\sum_{j=1}^k \left[\left(\sum_{i=1}^p \lambda_{ij} \right)^2 \right] + \sum_{i=1}^p s_i + \sum_{i=1}^p \epsilon_i}$$

Appendix B

```
# In order to generate multivariate normal data, load package MASS
library(MASS)
# in order to be able to replicate the outcome, choose a random seed
set.seed(2021).

## Setup
n <- 1000000          # number of individuals
ng <- 1              # number of general factors
nf <- 5              # number of group factors
nu <- 10             # number of specific (unique) factors
ne <- 10             # number of error terms
nv <- ng + nf + nu + ne # total number of (latent) variables

# generate values for the nv variables
# means are set to 0 for all variables
# sds to 1
# assume variables are all independent (diagonal variance-covariance matrix)
# empirical = TRUE means generate population data/'exact simulation'.
Data <- mvrnorm( n, rep( 0, nv ), diag( nv ), empirical = TRUE ).

# add column names to the dataset.
colnames( data ) <- c( 'g',

  paste0( 'f', 1:nf ),
  paste0( 'u', 1:nu ),
  paste0( 'e', 1:ne ))
```

Acknowledgments

I would like to thank Kees-Jan Kan for generating the data, Lauren McGrath for providing me with the reliabilities of the subtests of the US version of the WISC-V and Madelon van den Boer for her comments on an earlier version of this paper

Declaration of conflicting interests

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

Funding

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

ORCID iD

Peter F. de Jong  <https://orcid.org/0000-0002-8806-0563>

References

- Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63, 1–29. <https://doi.org/10.1146/annurev-psych-120710-100422>
- Beaujean, A. A. (2018). Simulating data for clinical research: A tutorial. *Journal of Psychoeducational Assessment*, 36(1), 7–20. <https://doi.org/10.1177/0734282917690302>

- Benson, N. F., Maki, K. E., Floyd, R. G., Eckert, T. L., Kranzler, J. H., & Fefer, S. A. (2020). A national survey of school psychologists' practices in identifying specific learning disabilities. *School Psychology, 35*(2), 146–157. <https://doi.org/10.1037/SPQ0000344>
- Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80*(4), 796–846. <https://doi.org/10.1111/j.1467-6494.2011.00749.x>
- Canivez, G. L., Dombrowski, S. C., & Watkins, M. W. (2018). Factor structure of the WISC-V in four standardization age groups: Exploratory and hierarchical factor analyses with the 16 primary and secondary subtests. *Psychology in the Schools, 55*(7), 741–769. <https://doi.org/10.1002/pits.22138>
- Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2016). Factor structure of the Wechsler Intelligence Scale for Children—Fifth edition: Exploratory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 28*(8), 975–986. <http://dx.doi.org/10.1037/PAS0000238>
- Canivez, G. L., Watkins, M. W., & Dombrowski, S. C. (2017). Structural validity of the Wechsler Intelligence Scale for Children—Fifth edition: Confirmatory factor analyses with the 16 primary and secondary subtests. *Psychological Assessment, 29*(4), 458–472. <http://dx.doi.org/10.1037/pas0000358>
- Canivez, G. L., & Youngstrom, E. A. (2019). Challenges to the Cattell-Horn-Carroll theory: Empirical, clinical, and policy implications. *Applied Measurement in Education, 32*(3), 232–248. <https://doi.org/10.1080/08957347.2019.1619562>
- Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—Simulation and Computation*[®], *39*(4), 860–864. <https://doi.org/10.1080/03610911003650383>
- Cronbach, L. J., & Furby, L. (1970). How we should measure “change”: Or should we? *Psychological Bulletin, 74*(1), 68–80. <https://doi.org/10.1037/h0029382>
- Dombrowski, S. C., Beaujean, A. A., McGill, R. J., Benson, N. F., & Schneider, W. J. (2019). Using exploratory bifactor analysis to understand the latent structure of multidimensional psychological measures: An example featuring the WISC-V. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(6), 847–860. <https://doi.org/10.1080/10705511.2019.1622421>
- Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2018). Factor structure of the 10 WISC-V primary subtests across four standardization age groups. *Contemporary School Psychology, 22*(1), 90–104. <http://dx.doi.org/10.1007/s40688-017-0125-2>
- Dombrowski, S. C., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2015). Exploratory bifactor analysis of the Wechsler Intelligence Scale for Children—fifth edition with the 16 primary and secondary subtests. *Intelligence, 53*, 194–201. <https://doi.org/10.1016/j.intell.2015.10.009>
- Dombrowski, S. C., McGill, R. J., & Morgan, G. B. (2021). Monte Carlo modeling of contemporary intelligence test (IQ) factor structure: Implications for IQ assessment, interpretation, and theory. *Assessment, 28*(3), 977–993. <https://doi.org/10.1177/1073191119869828>
- Dombrowski, S. C., McGill, R. J., Watkins, M. W., Canivez, G. L., Pritchard, A. E., & Jacobson, L. A. (2022). Will the real theoretical structure of the WISC-V please stand up? Implications for clinical interpretation. *Contemporary School Psychology, 26*(4), 492–503. <https://doi.org/10.1007/s40688-021-00365-6>
- Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., Frans, O., Gintiliene, G., Hagemester, C., Iliescu, D., Jaworowska, A., Jimenez, P., Manthouli, M., Matesic, K., Schittekatte, M., Sumer, H. C., Urbanek, T., & Halama, P. (2012). Testing practices in the 21st century. *European Psychologist, 17*(4), 300–319. <https://doi.org/10.1027/1016-9040/A000102>
- Farmer, R. L., McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2021). Why questionable assessment practices remain popular in school psychology: Instructional materials as pedagogic vehicles. *Canadian Journal of School Psychology, 36*(2), 98–114. <https://doi.org/10.1177/0829573520978111>
- Fletcher, J. M., & Miciak, J. (2017). Comprehensive cognitive assessments are not necessary for the identification and treatment of learning disabilities. *Archives of Clinical Neuropsychology, 32*(1), 2–7. <https://doi.org/10.1093/arclin/acw103>
- Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. John Wiley & Sons.

- Kranzler, J. H., Maki, K. E., Benson, N. F., Eckert, T. L., Floyd, R. G., & Fefer, S. A. (2020). How do school psychologists interpret intelligence tests for the identification of specific learning disabilities? *Contemporary School Psychology, 24*(4), 445–456. <https://doi.org/10.1007/s40688-020-00274-0>
- Lecerf, T., & Canivez, G. L. (2018). Complementary exploratory and confirmatory factor analyses of the French WISC–V: Analyses based on the standardization sample. *Psychological Assessment, 30*(6), 793–808. <http://dx.doi.org/10.1037/pas0000526>
- Macmann, G., & Barnett, D. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "Intelligent Testing" approach to the WISC-III. *School Psychology Quarterly, 12* (3), 197–234. <https://doi.org/10.1037/h0088959>
- McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8*(3), 290–302. <https://doi.org/10.1177/073428299000800307>
- McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology, 71*, 108–121. <https://doi.org/10.1016/j.jsp.2018>
- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. <https://doi.org/10.1016/j.intell.2008.08.004>
- Miciak, J., Taylor, W. P., Stuebing, K. K., & Fletcher, J. M. (2018). Simulation of LD identification accuracy using a pattern of processing strengths and weaknesses method with multiple measures. *Journal of Psychoeducational Assessment, 36*(1), 21–33. <https://doi.org/10.1177/0734282916683287>
- Pauls, F., & Daseking, M. (2021). Revisiting the factor structure of the German WISC-V for clinical interpretability: An exploratory and confirmatory approach on the 10 primary subtests. *Frontiers in Psychology, 12*, 710929. <https://doi.org/10.3389/fpsyg.2021.710929>
- Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders. *Cognition, 101*(2), 385–413. <https://doi.org/10.1016/j.cognition.2006.04.008>
- R Development Core Team. (2016). *R: A Language and Environment for Statistical Computing (Version 3.2.3) [Computer Program]*. R Foundation for Statistical Computing.
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696. <https://doi.org/10.1080/00273171.2012.715555>
- Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129–140. <https://doi.org/10.1080/00223891.2012.725437>
- Schneider, W. J. (2013). What if we took our models seriously? Estimating latent scores in individuals. *Journal of Psychoeducational Assessment, 31*(2), 186–201. <https://doi.org/10.1177/0734282913478046>
- Toffalini, E., Giofrè, D., & Cornoldi, C. (2017). Strengths and weaknesses in the intellectual profile of different subtypes of specific learning disorder: A study on 1, 049 diagnosed children. *Clinical Psychological Science, 5*(2), 402–409. <https://doi.org/10.1177/2167702616672038>
- van Bergen, E., van der Leij, A., & de Jong, P. F. (2014). The intergenerational multiple deficit model and the case of dyslexia. *Frontiers in Human Neuroscience, 8*, 346. <https://doi.org/10.3389/fnhum.2014.00346>
- van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842–861. <https://doi.org/10.1037/0033-295X.113.4.842>
- van der Maas, H. L. J., Savi, A. O., Hofman, A., Kan, K.-J., & Marsman, M. (2019). The network approach to general intelligence. In D. J. McFarland (Ed.), *General and specific mental abilities* (pp. 108–131). Cambridge Scholars Publishing.

- van Iterson, L., de Jong, P. F., & Zijlstra, B. J. (2015). Pediatric epilepsy and comorbid reading disorders, math disorders, or autism spectrum disorders: Impact of epilepsy on cognitive patterns. *Epilepsy & Behavior, 44*, 159–168. <http://dx.doi.org/10.1016/j.yebeh.2015.02.007>
- Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC–III subtest composite strengths and weaknesses. *Psychological Assessment, 16*(2), 133–138. <https://doi.org/10.1037/1040-3590.16.2.133>
- Watkins, M. W., & Canivez, G. L. (2022). Assessing the psychometric utility of IQ scores: A tutorial using the Wechsler intelligence scale for children–fifth edition. *School Psychology Review, 51*(5), 619–633. <https://doi.org/10.1080/2372966X.2020.1816804>
- Watkins, M. W., Canivez, G. L., Dombrowski, S. C., McGill, R. J., Pritchard, A. E., Holingue, C. B., & Jacobson, L. A. (2022). Long-term stability of Wechsler Intelligence Scale for Children–fifth edition scores in a clinical sample. *Applied Neuropsychology: Child, 11*(3), 1–7. <https://doi.org/10.1080/21622965.2021.1875827>
- Wechsler, D. (2014). *Wechsler Intelligence Scale for Children-Fifth Edition Technical and Interpretive Manual*. NCS Pearson.
- Wechsler, D. (2018). *Wechsler Intelligence Scale for Children-Fifth Edition Technical and Interpretive Manual Dutch Version*. NCS Pearson.