Informe técnico

Usos y abusos de las puntuaciones obtenidas a través de los test psicológicos

Pere Joan Ferrando (1), Urbano Lorenzo-Seva (1), Roberto Colom (2)

(1) Universitat Rovira i Virgili, (2) Universidad Autónoma de Madrid

Tarragona, 3 de marzo de 2022

WAIS

Participants: 1.369

Variables: 14 subtests (Vocabulary, Similarities, Arithmetic, Digit Span, Information, Comprehension, Letter_Number, Picture Completion, Coding, Block Design, Matrices, Picture arrangment, Symbol Search, y Object Assembly) del test WAIS III.

External criterion: Educational level (registered in 4 levels).

Basis analyses

The aim of this study is to determine the most appropriate scoring schema for the WAIS III measure. We want to decide between multiple scores derived from the correlated-factors solution previously identified, or general scores derived from an essentially unidimensional solution. This last solution, in turn, can be obtained by fitting a single factor (Spearman FA) or by fitting a second-order general factor obtained from the 4 correlated primary factors to the data.

The main difference between our study and the usual approaches for deciding which the most appropriate FA solution is, is that our decisions are not mainly based on the calibration and structural results (goodness of model-data fit, factorial structure) but on the properties of the factor score estimates derived from the factorial solutions. We consider this emphasis as justified: tests are measurement instruments, and so, the scores obtained from them are those that will be used for making individual assessments, predictions, or take clinical decisions.

In the line above, the decision about the "best" scoring for a particular instrument also depends on the use we want to give to the scores. If they are to be used for individual assessment, then the "internal" properties of accuracy (reliability) and determinacy are the most relevant. If they are used for prediction purposes, then predictive power becomes the most relevant decision criterion.

With this background, we start the comparative analyses. The first step is to assess the adequacy of the different solutions at the structural level. We start by considering the goodness-of-fit results:

Table 1

Goodness-of fit results

Model	RMSEA	CFI	GFI	AGFI	RMSR
1 factor	.118	.970	.994	.993	.0538
4 factors	.028	.999	1.000	.999	.0124
Second order		.976	1.000	.999	.0198

Note: RMSEA= Root Mean Square Error of Approximation; CFI=Comparative Fit Index; GFI=Goodness of Fit Index; AGFI=Adjusted Goodness of Fit Index; RMSR=Root Mean Square of Residuals

Comments: The GOF results allows us to anticipate further results. The correlated-factors model and the second-order model have an excellent fit in all the facets of fit considered. In fact, the fit is even too good, which clearly indicates that no more primary factors should be extracted from this data.

The fit of the single-factor model would not be considered acceptable using rigorous standards, and so, the usual approach in applications would be to discard the single-factor solution and accept the multiple or the second order solutions as the most appropriate. Things, however, are not so simple. The single-factor solution fits badly in relative terms (fit per degree of freedom) as evidenced by the RMSEA. However, the absolute fit in terms of GFI and RMSR is not too bad. And the comparative fit with respect to the null model of independence (CFI) is not bad either. These results anticipate that most of the common variance of the data will be accounted for by a single principal factor.

We shall now explore indices at the structural level that go beyond pure model data fit. Essential unidimensionality at the structural level can be assessed by considering (a) the results of parallel analysis (b) the amount of explained common variance, and (c) the strength and replicability of the single-factor solution. These results are:

Table 2

Indices at the structural level

(a) Parallel Analysis (PA). Based on minimum rank factor analysis

Variable	Real-data % of variance	Mean of random % of variance	95% percentile of random % of variance
1	70.4201*	14.6373	17.9827
2	6.9526	13.3037	15.6343
3	4.7471	12.0829	14.1638
4	3.5201	10.9391	12.5800

(b) Closeness to unidimensionality assessment

ECV	.924(.917;.935)**
MINREAL	.197(.177;.207)**

(c) Construct replicability: Generalized H (G-H) index.

Model		H-Latent
1 factor		.961(.957;.963)
4 factors	VC	.929(.921;.936)
	PO	.944(.938;.950)
	WM	.901(.889;.917)
	PS	.915(.903;.933)
Second-order solution		.961(.957;.963)

Note: ** BC bootstrap 95% Confidence Intervals; VC=Verbal Comprehension; PO=Perceptual Organization; WM=Working Memory; PS=Perceptual Speed

Comments: The additional information still leaves things undecided. The results of PA suggests that a strong general factor dominates the data. And, in agreement to this result, (a) the percentage of explained common variance that the first principal factor is able to explain is more than 90% and (b) the loadings on the second principal factor are rather low (MIREAL results). However, the generalized H results clearly suggest that the structure in four factors is strong and replicable for all of them. But note, that the H index is higher for the general or single factor than for any of the primary factors.

We turn now to the main focus of the comparison: the score properties. The first properties we shall assess are those concerned with the reliability, determinacy, and discriminating power of the factor score estimates derived from the different solutions

Table 3

Quality and effectiveness of factor score estimates.

Model		FDI	r_{xx} (ORION)	SR	EPDT
1 factor		.980	.961	4.938	96.1%
4 factors	VC	.964	.929	3.618	94.0%
	PO	.971	.944	4.093	94.9%
	WM	.949	.901	3.016	92.5%
	PS	.957	.915	3.283	93.2%
Second-order solution		.985	.971	5.736	96.9%

Note: VC=Verbal Comprehension; PO=Perceptual Organization; WM=Working Memory; PS=Perceptual Speed; FDI=Factor Determinacy Index; r_{xx}=marginal reliability (ORION approach); SR=sensitivity ratio; EPDT=Expected percentage of true differences.

Comments: The results in this section are impressive. In any of the scoring schemas, the score estimates are highly reliable, very determinate and with high discriminating power (any estimated score would correctly detect above 90% of true differences between trait levels). At the comparative level, the general scores are more accurate and discriminating than the primary scores. However, these results can be expected from basic psychometric principles, because the general scores can be regarded as linear composites of the primary scores.

In order to make a fair comparison and decide whether the primary score estimates use more information from the data than the general score estimates, we shall use the added-value approach: to assess whether the factor score estimates from a primary factor are more accurate predictors of the corresponding primary true scores than the score estimates from the general factor are. If they are, the primary scores have added value. The results are now.

Table 4

Added-value assessment

	PRMSE From the primary score estimates	PRMSE From the general score estimates	From both: primary factors and the General Factor
VC	.929(.921;.936)	.787	.963
PO	.944(.938;.949)	.865	.971
WM	.901(.888;.917)	.749	.948
PS	.915(.903;.933)	.773	.956

Note. PRMSE=Proportional reduction in mean squared error of prediction; VC=Verbal Comprehension; PO=Perceptual Organization; WM=Working Memory; PS=Perceptual Speed

The results are clear here. For each of the 4 primary factors, the 'true' levels in the factors are better predicted form the primary score estimates than from the general estimates. The conclusion here is that, **at the internal**

level, the use of general score estimates instead of primary score estimates entails a certain loss of information. Furthermore, this loss is significant if judged by the lower limits of the confidence intervals.

If the WAISS-III scores were to be used for accurate individual measurement or for taking clinical decisions, then the informed choice based on the internal analysis so far would be to use scores derived from the second order solution using both, the 4 primary scores and the general scores (see the last column above: the joint use of primary and general scores maximizes added value). This decision however, must be qualified. Thus, the use of general scores alone (either derived from the single factor or the second order solution) is well justified, as the solution for the WAISS-III can be considered to be essentially unidimensional. So, these scores have meaning and can be univocally interpreted, and, in addition to this, they have excellent internal properties. The only drawback is that simplicity is achieved at the cost of some loss of information.

But, what will happen if the interest is not in individual assessment but in predicting external outcomes?

To assess this issue, we shall use two sources of information as proposed by the authors: Differential validity and incremental validity. Differential validity assesses the extent to which the primary factor score estimates relate to the criterion in a different way as how they do relate to the general factor. If this condition is met, it can be inferred that some external validity information cannot be solely explained by the general factor. So, it is advantageous (in a validity sense) to consider the multiple factor model. If evidence of differential validity is not found, it can be inferred that all the external validity information can be "channeled" through the general factor. In this case, no further information is gained from the multiple model, and, on parsimony grounds, it is better to adopt the simpler essentially unidimensional solution.

The differential validity results are provided below. They are based on Bayes EAP score estimates:

Table 5.

Differential validity assessment

	$\hat{\rho}_{\hat{\theta}_{ky}} / \gamma_k$	90% CI
VC	.8448	(.8191 - .8740)
PO	.7096	(.6836 - .7375)
WM	.7620	(.7338 - .7971)
PS	.7271	(.6974 - .7585)

Note. CI=confidence interval; VC=Verbal Comprehension; PO=Perceptual Organization; WM=Working Memory; PS=Perceptual Speed

The results above suggest that some factors (particularly 1 and 2) are more strongly or weakly related to the criterion that can be predicted from their relations to the general factor. So, some validity information would be lost if an essentially unidimensional solution was considered in place of the 4-factor solution.

Incremental validity assessment involves comparing the predictive power of the general factor score estimates with respect to that of the 'best' composite (in the regression sense) of the primary score estimates. The basic idea here would be then to compare a simple correlation coefficient to a multiple correlation coefficient. However, the procedure proposed by the authors corrects for measurement error (which differentially affects both sources of evidence) and allows a fair comparison to be made. The results, based again on EAP estimates are below.

Table 6.

Incremental validity results

Incremental validity assessment	dif	Incremental value estimate
.6698(.6487;.6933) .7196(.7029; .7379)	.1003(.0728;.1277)	.0498(.0298;.0698)**

Note. dif = differential.

The results are consistent to those concerned to differential validity above. Once the corrections have been made, the predictive power of the optimal composite of the primary score estimates is non-negligibly and significantly better to that obtained from the simple general score estimates. In practical terms, however, the gains achieved with the multiple solution are small

Follow-up analyses

Extension1: The behavior of raw scores when used in place of factor score estimates

All the results in the basis analyses above are based on full information ORION EAP score estimates that can be considered as optimal in the sense of providing the estimates that are most accurate and closer to the true trait levels. The issue that we attempt to address in this first extension is to assess the loss of accuracy and information, both in internal and in external terms, that is expected to be lost if simple raw scores are used in place of the EAP score estimates.

Internal analysis. Reliability assessment

Table 7

Reliability estimates for the EAP-ORION scores and the raw scores in the selected solutions.

Model		r_{xx} (ORION)	r_{xx} (Alpha)
4 factors	VC	.929	.869
	PO	.944	.763
	WM	.901	.860
	PS	.915	.747
Second-order solution		.971	.912

Note: VC=Verbal Comprehension; PO=Perceptual Organization; WM=Working Memory; PS=Perceptual Speed; FDI=Factor Determinacy Index; r_{xx} (ORION)=marginal reliability of the EAP estimates (ORION approach); r_{xx} (Alpha)=marginal reliability of the raw scores (Alpha estimate);

Comments. The reliability estimates of the raw scores are acceptable by most stablished standards but clearly lower than those of the EAP scores. Note also that the loss of accuracy is not homogeneous, but is more marked in the PO and PS factors. A possible explanation for this differential decreasing is in the "borrowing strength" phenomenon. The ORION scores use the information provided by the inter-factor correlations (which are rather high). The reliability of the raw scores is estimated in a separate scale-by-scale basis and does not make use of this information.

A second index for assessing the relative quality of the two scoring schemas is the coefficient of fidelity, which assesses the correlation between the scores at hand and the 'true' levels in the factor they attempt to measure. The results for this coefficient are

Table 7

Fidelity coefficients for the EAP-ORION scores and the raw scores in the selected solutions.

Model		Fidelity (ORION)	Fidelity(Raw)
4 factors	VC	.964	.954
	PO	.971	.960
	WM	.949	.932
	PS	.957	.936
Second-order solution		.985	.978

Comments: The results in table 7 suggests that the raw scores are good proxies for the factor scores they represent. So, the loss caused by the use of the simpler raw scores appears to be more on accuracy (reliability) than on factor representativity.

Finally, let us to assess the behavior of the raw scores in relative validity terms. First we obtain the product moment correlation between the criterion on the one hand, and (a) the general factor score estimates, and (b) the raw total scores.

Table 8

Correlations between the criterion and the two unidimensional score estimates

	Factor score estimates	Raw total scores
Criterion	.694	.673

Comments: The factor score estimates are better predictors, as expected. They are more reliable and closer to the 'true' levels. The difference in 'brute' predictive power, however is admittedly small.

We now report the corresponding correlations on a factor-by-factor (scale by scale) basis. They are:

Table 9

Correlations between the criterion and (a) EAP-ORION scores and (b) the raw scores on a factor-by-factor basis

	EAP (ORION)	Raw
VC	.694	.694
PO	.655	.609
WM	.636	.603
PS	.624	.568

Comments: Except for the first factor, in the remaining cases the EAP-ORION estimates perform better than the raw scores in validity terms. This profile suggests again a "borrowing strength" phenomenon. The strongest primary factor is VC and there are not validity differences here. In the remaining cases, ORION uses the auxiliary information provided by the correlations between the primary factors.

We turn now to the multiple correlation analyses. The multiple correlation coefficients between the criterion and (a) the primary factor score estimates, and (b) the raw sub-scale scores are:

Table 10

Multiple Correlation between the criterion and the primary factor score estimates (EAP-ORION) and the subscale scores

	Factor score estimates	Subscale scores
Criterion	.710	.704

Comments: In multiple-correlation terms the ORION estimates attain a higher validity estimate, but the difference is small. Compared to the simple criterion-total correlations the multiple correlations are slightly

higher, as the Univalue-added-value analyses suggested previously. In practical terms, however, the differences are small.

Extension2: Person-fit analysis

So far, we have made a proposal regarding (a) the most appropriate structure for the WAIS-C and, above all, (b) the most appropriate scoring schema for measuring this structure the best as possible. However, it cannot be uncritically assumed that the best choices, that are based on the responses at the general group level, apply to each of the individuals that responded to the test. In other words, the best scoring schema we propose for the whole group, is still compatible with the presence of a certain proportion (presumably small) of respondents that do not fit the model. As a consequence, the scores of these missfitting individuals cannot be meaningfully and validly interpreted. In the worst scenario, the score of a missfitting individual can be totally meaningless.

In order to detect potential respondents whose scores are not in agreement with the model, we have used two simple indices that work well in practice (ref Frontiers). The first is a model-based chi-square type-weighted residual statistic, the WMSI, which assesses the discrepancy between the vector of responses provided by the individual, and the responses that would be expected given the chosen model and the score estimate of this individual. The second index is a group-based (not model based) correlational-type statistic: the personal correlation, which is the product moment correlation between the vector of responses of the individual and the vector of average responses in the group that was analyzed.

The results below show the most missfiting respondents in the group as flagged by both statistics.

Table 11
Sorted by WMSI

Case	WMSI	Rp
153	7,282	.914
546	6,913	.185
1128	6,877	.321
461	5,76	.834
481	5,313	.825
1265	5,183	.922
1205	4,647	.888
1356	4,518	.685
54	4,515	.802
1054	4,152	.916
701	4,131	.897
67	4,011	.911
535	3,997	.924
1081	3,981	.942
200	3,978	.799
885	3,921	.926
114	3,91	.941
577	3,803	.859
109	3,781	.866
852	3,578	.769
1049	3,573	.914
349	3,555	.904
203	3,538	.836
62	3,486	.958
3	3,401	.844
1011	3,371	.637
1226	3,344	.943
993	3,31	.718
1329	3,275	.856
939	3,241	.625
1005	3,122	.935
1125	3,08	.512
1050	3,071	.941
616	2,981	.881

Table 12
Sorted by personal correlation

Case	WMSI	Rp
292	.751	.036
449	1.567	.092
55	.931	.101
398	.581	.114
206	1.145	.158
125	.698	.162
18	.466	.173
120	1.456	.18
546**	6.913	.185
278	.527	.229
44	.551	.249
249	. 343	.303
1128**	6.877	.321
730**	2.261	.33
51	.268	.336
79	.596	.336
16**	2.109	.341
396	.257	.341
565	.582	.352
606	1.385	.371
128	.743	.378
561	1.169	.378
285	.603	.401

Of particular relevance in the list above are those individuals detected as potentially inconsistent by both indicators (doble *). Thus, for example, respondent no 546 has a very high WMSI value (FACTOR cut-off value is 2) and a very low personal correlation. The first result might be obtained by unexpectedly low or high scores in certain of the 14 parcels or subtests. The second, however, suggests that the score profile of this individual across the 14 subtests has a different shape than the consensus group profile. We note in closing that EAP scores are provided for all the participants in the group. However, valid interpretation of the estimated scores for the individuals doubly-flagged by the person-fit measures cannot be warranted. Finally, is spite of the very high discrepant values above, the person-fit results suggest that most of the participants responded to the WAIS in a rather consistent way.

TASC

Participants: 1.022

Variables: Anxiety Scale For Children (TASC) (Sarason, et al, 1960) has 30 binary items.

1. I wonder if I will pass
2. My heart beats fast
3. I look around the room
4. I feel nervous
5. I think I am going to get a bad grade
6. It is hard for me to remember the answers
7. I play with my pencil
8. My face feels hot
9. I worry about failing
10. My belly feels funny
11. I worry about doing something wrong
12. I check the time
13. I think about what my grade will be
14. I find it hard to sit still
15. I wonder if my answers are right
16. I think that I should have studied more
17. My head hurts
18. I look at other people
19. I think most of my answers are wrong
20. I feel warm
21. I worry about how hard the test is
22. I try to finish up fast
23. My hand shakes
24. I think about what will happen if I fail
25. I have to go to the bathroom
26. I tap my feet
27. I think about how poorly I am doing
28. I feel scared
29. I worry about what my parents will say
30. I stare

External criterion: Academic performance in Mathematics and Neuroticism.

Preliminary notes: Unlike the previous example, the units of the analysis here are individual items, and, furthermore, the items are binary. Given this scenario, the data will be fitted here using a non-linear FA model based on the Underlying-variables-approach, and the score estimates will no longer be linear composites of the observed scores but have to be obtained iteratively. The resulting estimates are non-linearly related to the raw scores.

Basis analyses

We want to decide here between a two-factor solution with correlated factors or an essentially unidimensional solution. For starting, the goodness of fit results are:

Table 13
Goodness-of fit results

Model	RMSEA	CFI	GFI	AGFI	RMSR
1 factor	.058	.975	.961	.961	.0792
2 factors	.047	.985	.974	.970	.0644

Note: RMSEA= Root Mean Square Error of Approximation; CFI=Comparative Fit Index; GFI=Goodness of Fit Index; AGFI=Adjusted Goodness of Fit Index; RMSR=Root Mean Square of Residuals

Again, the fit of the single-factor model would not be considered acceptable by using rigorous standards, but is not too bad either and none of the indicators give clearly unacceptable values. The fit of the two-factor solution is good. There is no need to fitting additional factors.

We turn now to the indices for assessing essential unidimensionality (a) the results of parallel analysis (b) the amount of explained common variance, and (c) the strength and replicability of the single-factor solution. These results are:

Table 14
Indices for assessing essential unidimensionality

(a) Parallel Analysis (PA). Based on minimum rank factor analysis

Variable	Real-data % of variance	Mean of random % of variance	95% percentile of random % of variance
1	38.3314*	6.6957	7.2811
2	6.9831*	6.3024	6.7933
3	5.5379	6.0043	6.4411
4	4.8188	5.7439	6.1049

(b) Closeness to unidimensionality assessment

ECV	.853 (.829;.893)**
MIREAL	.201 (.164;.220)**

(c) Construct replicability: Generalized H (G-H) index.

Model		H-Latent
1 factor		.948 (.942;.953)
2 factors	F1	.905 (.881;.913)
	F2	.942 (.929;.951)

Note: *Dimensions to retain; ** BC bootstrap 95% Confidence Intervals; F1: Psychological Anxiety; F2: Physiological Anxiety.

Comments: The additional information start to tilt thing towards a general factor. The results of PA suggests that a strong general factor dominates the data; the percentage of explained common variance that the first principal factor is able to explain is above 85% with low loadings in the second principal factor, and, finally the H index is higher in the general factor than in any of the two factors.

In closing the structural part, we show the estimated factor pattern in the bidimensional solution together with the inter-factor correlation matrix. Note that the correlation between factor 1 and factor 2 is rather high

Table 15
Rotated loading matrix

Variable	F 1	F 2
V 1	.387	
V 2		.307
V 3		
V 4	.729	
V 5		
V 6	.306	.327
V 7	.890	
V 8		.520
V 9	.377	
V 10	.723	
V 11		.518
V 12	.388	.398
V 13	-.619	.732
V 14	.556	
V 15	.395	.399
V 16		.421
V 17		.605
V 18	.647	
V 19		.716
V 20	-.324	1.012
V 21		.674
V 22		.410
V 23		.727
V 24		.569
V 25		.743
V 26		
V 27		.577
V 28		.976
V 29	.655	
V 30		.580

Notes: Inter-factor correlation: .772; F1: Psychological Anxiety; F2: Physiological Anxiety;

We turn now to the score properties starting with the reliability, determinacy, and discriminating power of the factor score estimates derived from the competing solutions.

Table 16
Quality and effectiveness of factor score estimates.

Model		FDI	r_{xx} (ORION)	SR	EPDT
1 factor		.974	.948	4.270	95.2%
2 factors	F1	.873	.763	1.794	89.7%
	F2	.915	.837	2.269	91.8%

Note: FDI=Factor Determinacy Index; r_{xx}=marginal reliability (ORION approach); SR=sensitivity ratio; EPDT=Expected percentage of true differences; F1: Psychological Anxiety; F2: Physiological Anxiety.

Comments: There is less uncertainty here than in the previous example. Compared to the primary score estimates, the unidimensional score estimates are clearly more reliable, determinate, and able to provide effective discrimination among individuals. Let us see whether the added-value analysis provide more support to this trend.

Table 17
Added-value assessment

	PRMSE From the primary score estimates	PRMSE From the general score estimates	From both: primary factors and the General Factor
F1	.763 (.642; .806)	.678	.866
F2	.837 (.718; .875)	.758	.911

Note. PRMSE=Proportional reduction in mean squared error of prediction; F1: Psychological Anxiety; F2: Physiological Anxiety.

Comments: So as to conclude that the primary scores have added value, the lowest end of the confidence interval around their MSE-reduction estimate should be higher that the estimate obtained from the single general factor. This is not the case, neither for F1 nor for F2. The conclusion is that, **at the internal level**, there is no loss of information if the general score estimates are used instead of a multiple primary scores.

We turn now to assessing the behavior of the competing scoring schemas in terms of predicting external outcomes.

Table 18.
Differential validity assessment with marks in mathematics as external criterion

	$\hat{\rho}_{\hat{\theta}_{ky}} / \gamma_k$	90% CI
F1	-.1932	(-.2436;-.1343)
F2	-.4115	(-.4578;-.3579)

Note. F1: Psychological Anxiety; F2: Physiological Anxiety.

Table 17.
Incremental validity results with marks in mathematics as external criterion

Incremental validity assessment	Dif	Incremental value estimate
.4881(.4411;.5336) .3428(.2792;.3951)	.1091(.0583;.1583)	.1453(.1054;.1888)

Note. dif = differential.

Table 19.
Differential validity assessment with score in Neuroticism as external criterion

	$\hat{\rho}_{\hat{\theta}_{ky}} / \gamma_k$	90% CI
F1	.4724	(.4236;.5170)
F2	.5370	(.4894;.5825)

Note. F1: Psychological Anxiety; F2: Physiological Anxiety.

Table 19.
Incremental validity results with score in Neuroticisma as external criterion

Incremental validity assessment	Dif	Incremental value estimate
.5848 (.5313;.6362) .5757 (.5213;.6270)	.0323 (-.0143;.0788)	.0091(.0020;.0165)

Note. dif = differential.

Comments. Results can be summarized as follows. For the math criterion, the results suggests some evidence of differential and incremental validity although the effects are rather small. For the N criterion, however, there is no evidence at all of differential or incremental effects, thus suggesting that all the relations between the primary score estimates and the N scores are fully "channeled" through the general factor.

Conclusion from the basis analyses: In this case, the unidimensional solution would be, overall, the most appropriate choice

Follow-up analyses

Extension1: The behavior of raw scores when used in place of factor score estimates

As in the previous study, we attempt to address the eventual loss of accuracy and information, both in internal and in external terms, that is expected to be observed if simple raw scores are used in place of the EAP score estimates. This issue is more relevant here than in the previous example, because we have used the non-linear-UVA FA as a calibration model, and so, the raw scores and the EAP estimates are nonlinearly related. Because the basis analyses has lead us to choose the unidimensional solution as the most appropriate, we shall only focus on the results concerned with this solution

Internal analysis. Reliability assessment

Table 20

Reliability estimates for the EAP-ORION scores and the raw scores in the selected solution.

Model	r_{xx} (ORION)	r_{xx} (Alpha)
General Factor	.948	.890

Note:; r_{xx} (ORION)=marginal reliability of the EAP estimates (ORION approach); r_{xx} (Alpha)=marginal reliability of the raw scores (Alpha estimate);

Comments. In terms of the chosen unidimensional solution, the reliability estimate of the raw scores is quite acceptable by most stablished standards but clearly lower than the excellent estimate of the EAP scores.

A second index for assessing the relative quality of the two scoring schemas is the coefficient of fidelity, which assesses the correlation between the scores at hand and the 'true' levels in the factor they attempt to measure. The results for this coefficient are

Table 21

Fidelity coefficients for the EAP-ORION scores and the raw scores in the selected solution.

Model	ORION	Raw
General Factor	.974	.920

Comments: The results suggests that the raw scores are good proxies for the factor scores they represent. As in the previous study, the loss caused by the use of the simpler raw scores appears to be more on accuracy (reliability) than on factor representativity.

Finally, let us to assess the behavior of the raw scores in relative validity terms. To do so, we obtain the product moment correlation between the criteria on the one hand, and (a) the general factor score estimates, and (b) the raw total scores.

Table 22

Correlations between criterion 1 (maths) and the two unidimensional score estimates

	Factor score estimates	Raw total scores
Criterion	-.303	-.290

Table 23

Correlations between criterion 2 (N) and the two unidimensional score estimates

	Factor score estimates	Raw total scores
Criterion	.510	.500

Comments: In both cases, the factor score estimates reach larger predictive values. However, the differences are rather small and virtually negligible in practice.

Extension2: Person-fit analysis

The same indices used in the WAIS examples would be used here for detecting potentially inconsistent respondents whose scores might not be validly interpreted. In this case, however, the inconsistencies are assessed on an item by item basis.

Table 24
Sorted by WMSI

Case	WMSI	rp
2	1,713	-.16
960	1,687	-.13
498	1,646	-.109
1	1,62	-.073
427	1,62	-.052
914	1,556	-.002
512	1,543	-.162
957	1,542	.041
876	1,536	.023
920	1,528	.026
578	1,527	.054
143	1,517	-.005
56	1,513	.016
727	1,509	.01
1018	1,489	.084
549	1,482	.093
410	1,47	-.03
619	1,464	.094
811	1,462	.074

Table 25
Sorted by rp

Case	WMSI	rp
680	1,145	-.431
786	.666	-.248
782	.621	-.243
512	1,543	-.162
2	1,713	-.16
944	.953	-.159
783	.585	-.156
923	.702	-.15
784	.623	-.143
925	.93	-.137
960	1,687	-.13
857	.474	-.121
634	.659	-.117
384	.832	-.115
329	.739	-.114
498	1,646	-.109
671	.67	-.096

In general the values are substantially lower than in the previous example. However, this result possibly reflects that the indices (intended for continuous responses) are less powerful in the case of binary responses. We believe that, in this case, the personal correlation is more informative. Near zero values of this statistic suggests that the pattern of responses of this child is totally insensitive to the relative difficulty of the items, and so that the response behavior is possibly quite at random. A negative correlation suggests that the response pattern is opposite to the normative pattern. So, the respondent tends to endorse (agree) the most 'difficult' items and not to endorse the easier ones. This behavior might indicate misunderstanding of the test instructions or even sabotaging. In closing, we note that the majority of respondents appear to have answered the TASC in a rather consistent way.