Høgskolen
i Innlandet

Jo Kleiven, Per Normann Andersen og
Ulrika Håkansson

The «Yellow-Red test»

A closer look at data from a longitudinal pilot project with Norwegian children

Skriftserien 19 - 2022

Utgivelsessted: Elverum

Forfatteren er selv ansvarlig for sine konklusjoner.
Innholdet gir derfor ikke nødvendigvis uttrykk for høgskolens syn.

I Høgskolen i Innlandets skriftserie publiseres både internt og eksternt finansierte FoU-arbeider.

ISSN: 2535-5678
ISBN trykt utgave: 978-82-8380-362-4
ISBN digital utgave: 978-82-8380-363-1

Sammendrag

Hensikten med denne rapporten har vært å finne fram til interessante spørsmål for det videre arbeidet med testbatteriet Yellow/Red. Ganske enkle analyser har vært utført, delvis på grunn av datautvalgets begrensede størrelse (N=148).

Testskårene forbedres over tid, i likhet med andelen *riktige* responser. De ulike testleddene eller oppgaven i hvert spill gir ganske ulike responser, og disse forskjellene er forholdsvis konstante over flere gjentatte målinger. De originale summerte skårene synes likevel å gi rom for forbedringer både i reliabilitet og validitet.

Noen alternative kodingsmuligheter blir derfor antydet, men blir hverken bekreftet eller avkreftet av våre forsøk på rekoding. Det er likevel mulig at noen av omkodingsidéene kan bli utprøvd på en bedre måte med et større utvalg.

Emneord:

Eksekutive funksjoner hos barn, testmetoder

Oppdragsgiver:

Oppland fylkeskommune

Abstract

The intention of this report has been to identify interesting questions for further work on the Yellow/Red test battery. Rather simple analyses have been carried out, partly due to the limited size of the sample (N=148).

The test scores consistently improve over time, as does the percentage of *correct* responses. Also, items within the same game elicit rather different responses, and item differences remain rather constant across replications. However, the original summed scales may have room for some improvement in both reliability and validity.

Consequently, alternate coding paradigms are suggested. They are neither proven nor disproven, however, by our initial recoding attempts. With larger samples, the recoding ideas may be put to a better test.

Keywords:

Childrens' executive functions, test methods

Financed by:

Oppland county

Preface

In the *Art of Learning* project (Andersen et al., 2019), psychologists in the college at Lillehammer are investigating possible effect of an art-based school program on children's executive functions. In a pretest, the *Yellow-Red* test (Rosas et al., 2020) was employed in its Norwegian version. A sample of normal school children were tested before, under and after exposure to diverse art experiences in the classroom; and was compared to children that were not subject to these experiences. The resulting longitudinal data set is rather comprehensive (N > 130). An account of the content and practicalities of the test is given by (Kleiven et al., 2022).

Further use of the test is planned not only in Norway, but also in other countries. Consequently, information about the test may not only be interesting to our Norwegian colleagues. A set of preliminary analyses of our available data, therefore, may be found in the present report, written in English.

The Chilean author of the test, Professor Ricardo Rosas, made it clear to us that the «Yellow-Red» may not yet have reached its final stage, as it still is under development. We thus of course hoped that our thoughts and discussions would be helpful to the continued efforts with the *Yellow Red* in Chile.

Apparently, our hopes are now proving realistic. Our colleagues in Chile have not only given helpful comments to earlier drafts of the present paper. They have also seriously considered some our questions and ideas, and relevant changes in the Yellow-Red scoring algorithms are now on their way (Rosas-Días et al., 2022). They will of course be implemented in our continued use of the Yellow-Red test in Norway.

Lillehammer, June 2022

Jo Kleiven Per Normann Andersen Ulrika Håkansson

CONTENTS

1. BACKGROUND

The executive functions most commonly discussed in the literature are inhibition, working memory, and cognitive flexibility (Lehto et al., 2003). The Android-based *Yellow/Red*-application was was intended to tap these functions in children. Parts of the test have proven useful in research (Cf., e.g. Rosas et al., 2017). It was developed by *Centro UC Tecnologias de Inclusión* at the Psychology department of the *Pontificia Universidad Católica de Chile*.

A Norwegian version was developed in cooperation with prof. Per Normann Andersen at the Inland Norway University of Applied Sciences. It was used in a pilot study of the development of learning in Norwegian children (Hundevadt & Klausen, 2019), focusing on the impact of art experiences in school on learning. *Yellow/Red* data were collected before, during, and after the children's art experiences in the classroom; and were available for the present analyses.

In an article about the pilot study (Andersen et al., 2019), the BRIEF inventory (Davidson et al., 2006) was employed to assess executive functions. Although the *Yellow/Red* test was also used, this part of the data was not reported. Nonetheless, data from the *Yellow/Red* as well as the *BRIEF* are available, and thus may be compared. Hopefully, this will be useful to the mutual validation of the two methods.

In the *Yellow-Red* procedure, a brief introduction comes first. It is followed by four different tests, all of which are designed as a data game for an Android tablet. Each game/test contains several tasks, which may be viewed as the items of the test. Rather different tasks are included in the four games, with the intention of tapping different executive functions or aspects. Table 1 displays the name of each game/test, its purpose (according to Rosas et al. (2020)), and the number of items included. More detailed accounts of these differences are given by Rosas et al. (2020) and (Kleiven et al., 2022).

Table 1: The four games of the Norwegian Yellow/Red test

Name of game	**Purpose**	**Number of items**
Cat/Dog	Exec. functions generally	33
Trios	Cognitive flexibility	21
Arrows	Inhibition	36
Bindings	Working memory	27

Initially, the response to each item were independently coded into a number. The coding rules were somewhat complex, and also different in the four games (Cf. (Rosas et al., 2020), (Håkansson et al., In preparation, 2022). The four different coding schemes are shown in figure 1 on the next page.

With the Cat/Dog and the Arrows games, correct responses are coded as +1, and errors as -2. A lacking response receives a neutral 0, while missing data are left out in all computations. The coding for Trios is rather similar, except for the introduction of a "Triple errors" category of responses. This special response receives the punishing score of -6. In all three games, therefore, positive scores indicate success with the item; while negative scores designate failures.

The Bindings game is different. Here, items have two different levels of difficulty. Correct responses to the easier level yield a score of +1; while success with a more difficult item is awarded +2 points. Perhaps more surprising, errors in this game receive a neutral 0 (not a negative number). Missing data are excluded from all further computations, just as they are in the three other games.

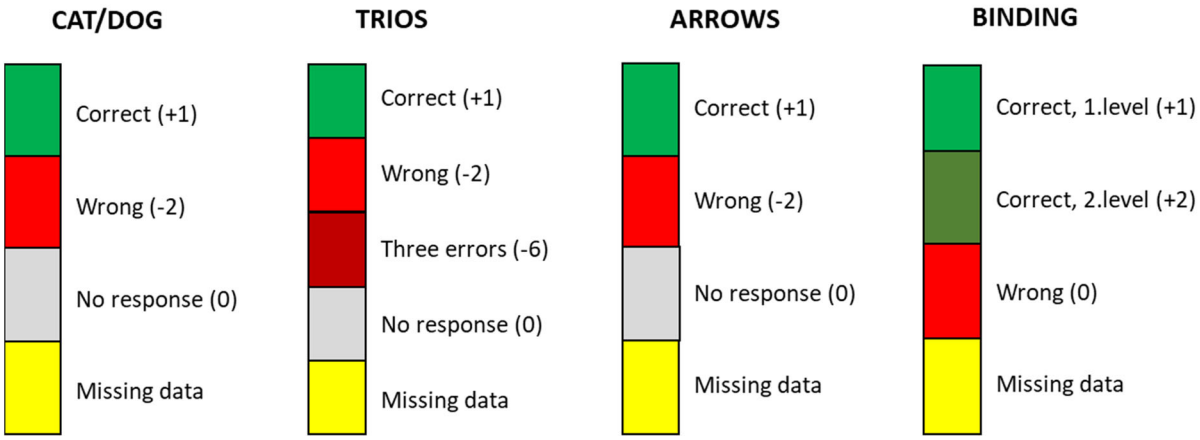

Figure 1: Recoding in the four games

Secondly, all codes for each game were summed into a 'scale' score for that game. In all scale scores, then, relative success in the game is shown by higher numbers. An opposite signal is given by low (or negative) summed scores, indicating more problems with the game.

The basic design of the project (Andersen et al., 2019; Hundevadt & Klausen, 2019) includes one experimental group and one control group. In the experimental group (N=83), diverse art experiences were used for classroom learning. These experiences were not available to the control group (N=56).

In addition, the design also had a longitudinal dimension. Both groups were tested

1. in January 2018 (before the intervention)
2. in May 2018 (directly after it), and
3. in September 2018 (six months later).

To keep things as simple as possible, only data from children participating at all three time points are used in our analyses (N=139). Listwise deletion of respondents, however, is not necessarily the best way of treating missing data (Allison, 2002). Other approaches should probably also be considered for more comprehensive analyses in the future.

2. THE MEASUREMENTS OF THE SUMMED SCALES

For a start, two 'classical' psychometric questions will be addressed. Are the summed measurements from the four games consistent or *reliable*? And to what extent are the four measures *valid*, measuring what they are intended to measure? Data from the three points in time will first be treated separately.

2.1. RELIABILITY

The assessment of reliability was not quite straightforward since the computation of *Cronbach's alpha* yielded a few surprises. With 'listwise deletion' of missing data, a substantial number of data points were lost. For several respondents, no response was recorded to one or more items. In the pre-intervention data set, especially, the number of dropouts was high in the Cat/Dog and Trio games. The corresponding *alpha* numbers are thus based on a reduced sample and may not be truly representative for the entire sample.

Table 2 displays the *alpha* values for the four games. As Pedhazur and Schmelkin (1991) have pointed out, there is no generally accepted standard for *alpha* values. For our present purposes, however, we view the alpha numbers of Cat/Dog and Arrows as acceptable. For Trios, however, the numbers need some improvement.

This problem is even more pressing with the Bindings game, however. Here, the final items are rather difficult; and the test is suspended after three incorrect responses. It is no surprise, therefore, that no one has responded to all items in any of the three points of time. Missing responses are most frequent with the five last tasks or items, where missing data are quite common, and no correct responses are given.

Unfortunately, *alpha* values for the complete set of items may then not be computed. Removing the five last items from the analysis also does not help. Attempting this, the mean covariance between the items turns *negative* ('negative average covariance among items'). This violates, of course, the statistical assumptions of the analysis. At any rate, it indicates that our measurements of the scale for this game are neither consistent nor reliable.

Table 2: Cronbach's *alpha* (and N) for data from four games at three time points

	Pre-intervention	**Just after int.**	**Six months later**
Cat/Dog (33 items)	0.87 (N = 18)	0.80 (N =103)	0.73 (N = 117)
Trios (21 items)	0.49 (N = 50)	0.49 (N = 83)	0.36 (N = 94)
Arrows (36 items)	0.69 (N = 122)	0.76 (N = 128)	0.68 (N = 125)
Bindings (27 items)	* (N = 0)	* (N = 0)	* (N = 0)

The *alpha* figures in the table are hardly compatible with the reliabilities reported by Garolera (2019), however. This apparent discrepancy may deserve some attention.

Even if data for some items are missing, however, a total score for each game may be computed for all individuals. Consequently, four test scores may be summed into a 'superordinate' score for the entire Yellow/Red test battery.

As shown by table 3, the sum of raw scores[1] yield a relatively weak *alpha* (0.62 - 0.63) for alle three time points, indicating that the *internal consistency* of the combined Yellow-Red scale is less than satisfactory.

The analyses perhaps also suggest that the main contribution to the problem may be the Bindings scale. At all points in time *alpha* is slightly decreased if the Cat/Dog, Trios, or Arrows scales are excluded. Excluding the Bindings scale, however, does not lead to reduced *alpha* values.

Table 3: Cronbach's alpha for *the sum of all Yellow/Red scales* at three time points

	Pre-intervention	**Just after int.**	**Six months later**
Yellow/Red summed	**.62**	**.63**	**.63**
Excluding Cat/Dog	.52	.55	.51
Excluding Trios	.53	.57	.57
Excluding Arrows	.46	.50	.47
Excluding Bindings	.63	.62	.64

2.2. VALIDITY

Within conventional psychometrics, assessing the *validity* of the Yellow/Red scales is even harder. As shown in table 1, the scales are intended to measure partly different aspects of executive functions. Ideally, then, some independent information on these characteristics or properties would serve as a standard for assessing the validity the variables. But the Yellow/Red scales are the only data on executive functions in our material, except for teacher responses to the BRIEF inventory (Gioia et al., 2000).

This is unfortunate. The YR scales and the BRIEF relate to rather different test conditions (Toplak et al., 2009), and thus may not be really compatible. There are also measurement problems with the BRIEF data.[2] Nonetheless, the BRIEF data is our only possibility for validating the Yellow/Red measures.

The eleven indices of the BRIEF all measure the *degree of problems* with a certain function. Consequently, high BRIEF scores mean more problems. The Yellow/Red scales, however, measure *the*

[1] Using z-scores or T-scores instead makes no great difference.

[2] More recently, this BRIEF version has been shown *not* to have satisfactory psychometric properties. Cf. Andersen, P. N., & Finbråten, H. S. (2020). Unsatisfactory psychometric properties of the Norwegian Behavior Rating Inventory of Executive Function Teacher Form - a Rasch Measurement Theory Validation. *Nevropsykologi*, *1*, 12-21.

degree of success. Higher numbers thus indicate relative success in the solving of tasks or items. If *less problems* in the BRIEF coincide with *more solved tasks* in the Yellow/Red, therefore, *significant* and *negative* correlations would be expected.

In table 4, pre-intervention data were used for the calculations. The limited number of negative and significant correlations are marked with yellow. Firstly, the Cat/Dog-scores are *only* correlated to Working Memory, and *not* to the ten remaining BRIEF indices. Since this game is intended to measure executive functions *generally*, the low number of significant relations is not quite as expected.

The Trios scale is also significantly related to the Working Memory index only, not to the others. The game's focus on cognitive flexibility thus is hardly supported.

In contrast, the Arrows scale yield several significant correlations. It covaries with Initiate, Working Memory and Organize Materials, as well as Metacognition and Global executive. This may perhaps be interpreted as a partial validation of the Arrows scale: It does at least measure parts of what is covered by some of the BRIEF indices. The scale's more precise intention to tap *cognitive flexibility* appears not to be validated by the BRIEF data, however.

The Bindings scale show no relation to any of the 11 BRIEF indices. Even here, then, the properties of the scale appear to be less than satisfactory. Normalizing the four summed scores and then combining them into an all-over score *(Suma de Z de pruebas)* also does not appear have more to offer. This general score is highly correlated only to indices *Initiate* (r = -0.285; p = 0.008), *Working Memory* (r = -0.266; p = 0.013), and *Metacognition* (r = -0.267; p = 0.022).

Table 4: Correlations between BRIEF-indices and 4 Yellow-Red scales, first data round

BRIEF index		Cat/Dog	Trios	Arrows	Bindings
Inhibit N=135	Pearson r	-0,09	-0,12	-0,10	-0,01
	Sig. (2-tailed)	0,27	0,15	0,25	0,89
Shift N=135	Pearson r	-0,04	-0,11	-0,12	-0,09
	Sig. (2-tailed)	0,67	0,20	0,17	0,30
Emotional control N=136	Pearson r	0,05	-0,10	-0,07	0,00
	Sig. (2-tailed)	0,53	0,25	0,40	0,99
Initiate N=136	Pearson r	-0,09	-0,14	-0,27	-0,13
	Sig. (2-tailed)	0,28	0,11	0,00	0,14
Working memory N=134	Pearson r	-0,19	-0,23	-0,31	-0,12
	Sig. (2-tailed)	0,03	0,01	0,00	0,17
Plan/organize N=136	Pearson r	-0,07	-0,09	-0,12	-0,11
	Sig. (2-tailed)	0,41	0,31	0,18	0,22
Organize materials N=126	Pearson r	-0,11	-0,14	-0,19	-0,04
	Sig. (2-tailed)	0,22	0,11	0,04	0,67
Monitor N=138	Pearson r	-0,06	-0,13	-0,12	-0,06
	Sig. (2-tailed)	0,49	0,12	0,17	0,52
Behavior regulation N=130	Pearson r	-0,04	-0,13	-0,11	-0,04
	Sig. (2-tailed)	0,66	0,14	0,20	0,66
Metacognition N=120	Pearson r	-0,12	-0,18	-0,26	-0,12
	Sig. (2-tailed)	0,20	0,06	0,00	0,20
Global Executive N=112	Pearson r	-0,07	-0,16	-0,20	-0,10
	Sig. (2-tailed)	0,46	0,09	0,04	0,31

A central question, then, is whether these relations between Yellow/Red and BRIEF are stable over time or not. Will the same picture emerge in the next two data sets? Table 5 on the following page displays the relevant correlations from the second round of data.

By and large, this is rather similar to the picture of table 4. A minor exception occurs with the Cat/Dog scale, however, which now is *not* correlated with any of the BRIEF indices. In addition, Arrows now is related to the Monitor index, and not to Organize Materials. Nonetheless, changes from the first to the second data round may be viewed as minor.

Table 5: Correlations between BRIEF-indices and 4 Yellow-Red scales, second data round

BRIEF index		Cat/Dog	Trios	Arrows	Bindings
Inhibit N=120	Pearson r	0,01	-0,05	-0,16	-0,02
	Sig. (2-tailed)	0,87	0,57	0,09	0,86
Shift N=121	Pearson r	-0,03	-0,04	-0,09	-0,05
	Sig. (2-tailed)	0,78	0,66	0,34	0,55
Emotional control N=123	Pearson r	0,08	0,05	-0,04	-0,02
	Sig. (2-tailed)	0,41	0,61	0,64	0,83
Initiate N=124	Pearson r	-0,12	-0,16	-0,26	-0,09
	Sig. (2-tailed)	0,18	0,08	0,00	0,34
Working memory N=125	Pearson r	-0,11	-0,18	-0,32	-0,08
	Sig. (2-tailed)	0,22	0,05	0,00	0,38
Plan/organize N=121	Pearson r	-0,12	-0,04	-0,16	-0,01
	Sig. (2-tailed)	0,18	0,63	0,08	0,95
Organize materials N=120	Pearson r	0,00	0,07	-0,12	-0,03
	Sig. (2-tailed)	0,99	0,46	0,21	0,72
Monitor N=122	Pearson r	-0,02	-0,04	-0,18	-0,04
	Sig. (2-tailed)	0,86	0,66	0,05	0,63
Behavior regulation N=117	Pearson r	0,03	0,01	-0,12	-0,03
	Sig. (2-tailed)	0,73	0,95	0,21	0,78
Metacognition N=113	Pearson r	-0,09	-0,08	-0,25	-0,04
	Sig. (2-tailed)	0,34	0,39	0,01	0,64
Global Executive N=109	Pearson r	-0,03	-0,04	-0,20	-0,04
	Sig. (2-tailed)	0,77	0,71	0,04	0,69

Finally getting to the data from the third round, however, differences appear: Substantial changes are shown in table 6 on the following page. Here, Cat/Dog and Arrows are both correlated with most BRIEF indices. This means that in the final wave of data the two scales receive some validation from the BRIEF measures of executive functions. The Cat/Dog scale, intended to be a *general* measure of executive functions, covaries with nine out of the eleven BRIEF indices. It also comes very close with the Organize Materials index. And the Arrows scale, focusing on inhibition, is finally confirmed by the Inhibit index of the BRIEF. It also correlates with nine other indices, possibly because the process of inhibition is generally important.

Table 6: Correlations between BRIEF indices and 4 Yellow-Red scales, third data round

BRIEF index		Cat/Dog	Trios	Arrows	Bindings
Inhibit N=97	Pearson r	-0,21	-0,07	-0,32	-0,11
	Sig. (2-tailed)	0,04	0,53	0,00	0,31
Shift N=99	Pearson r	-0,21	-0,11	-0,30	-0,11
	Sig. (2-tailed)	0,04	0,27	0,00	0,28
Emotional control N=96	Pearson r	-0,14	-0,06	-0,22	-0,18
	Sig. (2-tailed)	0,17	0,57	0,03	0,08
Initiate N=97	Pearson r	-0,30	0,01	-0,27	-0,10
	Sig. (2-tailed)	0,00	0,92	0,01	0,33
Working memory N=99	Pearson r	-0,38	-0,11	-0,33	-0,14
	Sig. (2-tailed)	0,00	0,27	0,00	0,17
Plan/organize N=97	Pearson r	-0,32	0,06	-0,22	-0,09
	Sig. (2-tailed)	0,00	0,59	0,03	0,39
Organize materials N=98	Pearson r	-0,20	0,09	-0,10	-0,10
	Sig. (2-tailed)	0,05	0,37	0,32	0,33
Monitor N=99	Pearson r	-0,29	-0,07	-0,36	-0,16
	Sig. (2-tailed)	0,00	0,49	0,00	0,12
Behavior regulation N=93	Pearson r	-0,21	-0,08	-0,32	-0,16
	Sig. (2-tailed)	0,05	0,46	0,00	0,14
Metacognition N=92	Pearson r	-0,32	0,02	-0,31	-0,12
	Sig. (2-tailed)	0,00	0,85	0,00	0,27
Global Executive N=88	Pearson r	-0,28	-0,02	-0,34	-0,14
	Sig. (2-tailed)	0,01	0,88	0,00	0,20

2.3. PRELIMINARY ASSESSMENT

The measurement properties of the Yellow/Red scales are not convincingly supported by the conventional, initial analyses. The test items appear not to quite measure the same thing, and the consistency or coherence between the different tasks is insufficient. The values of the Cronbach *alpha* thus are lower than desired. Also, the general *reliability* of the scales is hardly supported by the BRIEF measures.

Quite likely, the large amount of 'missing data' from the coding algorithms contribute to the reliability problem. In paragraph 5.1, this point will receive further comments.

Another point deserving attention is the question of *validity*: Do the scales actually measure executive functions as expected? This question is even more difficult to assess, however. The BRIEF inventory is the pilot project's only alternate source of information on the respondents' executive functions.

In tables 4 and 5 above, the correspondence between the Yellow-Red and the BRIEF data is neither tight nor convincing. Certain correlations do exist: BRIEF's Working memory index, e.g., shows some relationship to two of the four Yellow-Red scales. In addition, the Arrows scale is significantly correlated to several BRIEF-indices. Small and insignificant correlations, however, are the most common in the two tables by far.

But the figures of table 6 tell a different story. Most BRIEF indices correspond closely to the Cat/Dag and the Arrows scales, suggesting that the two methods partly measure the same phenomena. These two Yellow-Red scales thus are partly validated by the BRIEF indices. This is *not* so for the Trios and the Bindings scales, however. They are rather strikingly independent of the BRIEF indices and their information on executive functions.

It may also be noted, however, that the number of respondents (N) clearly is reduced in the final data round. An interesting question, therefore, is if this dropout influenced the correlations. Is it possible, e.g., that the children that remained in the final round of data collection yielded more clear data with less 'noise' than the dropouts had given in the two previous rounds?

At any rate, there may be several reasons why the Yellow-Red and the BRIEF give different information. As previously mentioned, the relatively low correlation between YR and BRIEF may reflect different test conditions (Toplak et al., 2009). YR seeks to measure underlying EF capabilities (cognition) in a well-structured situation, while BRIEF measures EF behaviour in everyday settings. Standard laboratory EF tasks lack ecological validity as context and modalities have relatively little in common with everyday behavioural demands (McAuley et al., 2010). This might explain why BRIEF is better correlated with academic performance than neuropsychological measures of EF (Pino Munoz & Aran Filipetti, 2019), such as YR.

In view of this complexity, only tentative and cautious conclusions are in order. Our preliminary analyses neither prove nor disprove the measurement properties of the Yellow Red scales. Hopefully, however, questions of reliability and validity will be addressed in better ways in the continued development of the scales.

3. LONGITUDINAL CHANGES

Of course, data from three different time points need to be compared. Are our measurements, e.g., stable over time, or do they change?

For each point in time, three versions of the summed scores exist for all four game scales (Cat/Dog, Trios, Arrows and Bindings). The first version is the apparently simple summed scores for each scale. In the second version, a conversion to *z* scores has taken place, and the third contains a *T* score conversion. In addition, each respondent's four z scores are summed into the collected index «*Suma de Z de pruebas*» (The sum of the tests). This index is also converted to the T score variable named «Global Index».

All versions, therefore, are converted from the four initial summed scores. For a start, therefore, the focus will be on the original Cat/Dog, Trios, Arrows and Binding scores for all respondents. Secondly, the collected index of «*Suma de Z de pruebas*» will be discussed.

3.1 THE CAT/DOG SCALE

This game has 33 items or tasks, and summed scores between 0 and 33 might be a reasonable expectation. As we have mentioned earlier, however, this is *not* a simple summed score, since the 'punishment score' of -2 for errors is used. The scores for the first time point, therefore, fall between -21 and +30, with a mean of 6,6. As shown in Figure 2, the distribution of scores comes close to having the familiar bell-shaped form.

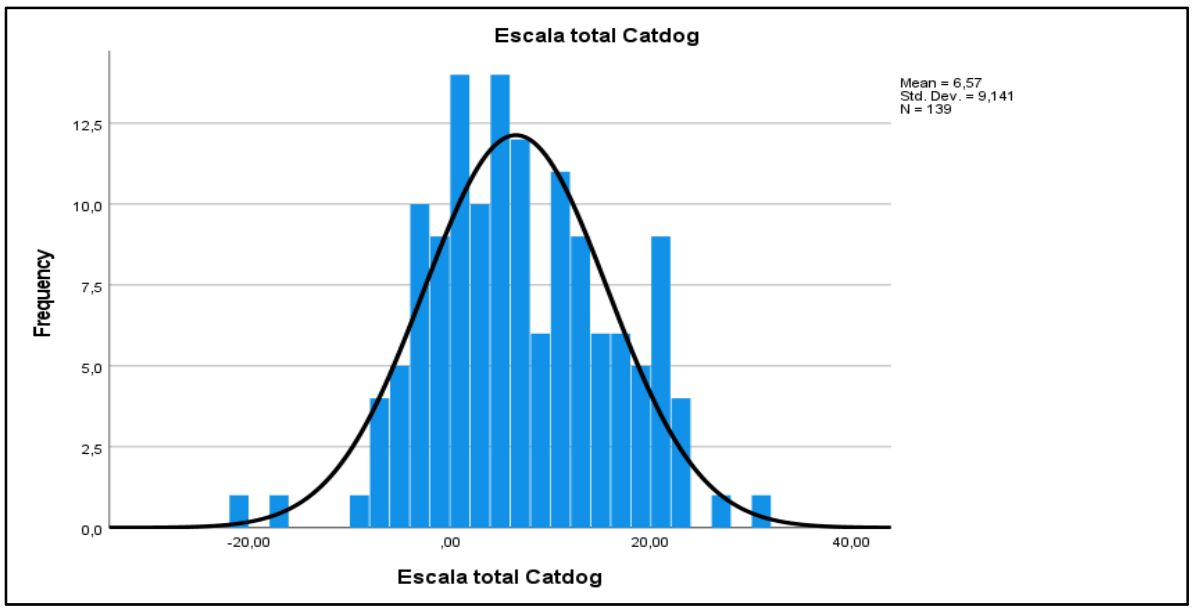

Figure 2: Distribution of the Cat/Dog variable at the first time point (pre-intervention)

Similar, bell-shaped distributions also occur for the data from the two remaining time points. There, means were 12,7 and 16,7; respectively. Corresponding maximal and minimal values were -16/+32 and -15/+32.

Figure 3 displays the mean at the three time points for the intervention and control groups. The means for the control group are consistently higher than those of the intervention group. This difference is also confirmed by ANOVA (MS = 1483,875; F = 8,241; p = 0,005). Does this indicate an unfortunate drawing of group members? Or may there be other explanations for this difference?

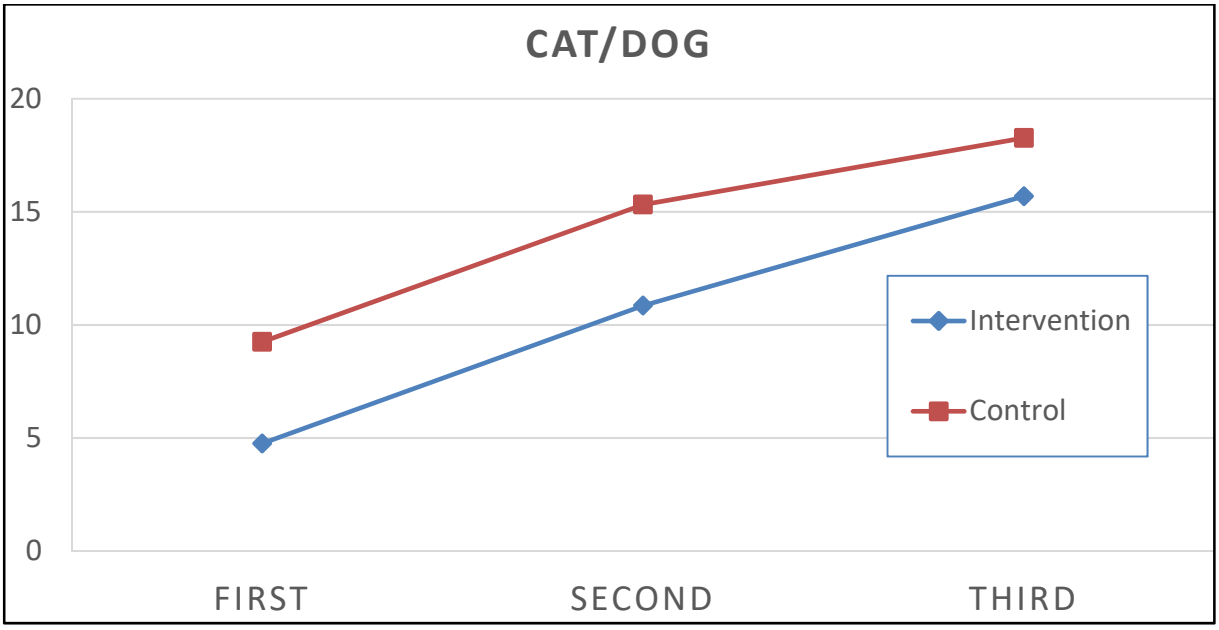

Figure 3: Group means of the Cat/Dog variable at three time points

The figure also shows that the means consistently increase, from the first through the second to the third wave of data. This also is confirmed by ANOVA (MS = 3379,413; F = 75,506; p < 0,001). Is this due to maturation or growth in the children, or to learning? Or may there be other viable explanations?

In the «Learning to Learn» project, one might hope for the mean increase to be higher in the intervention than in the control group. In that case, an interaction effect would appear between the «repeat» and the «group» factors. No such effect may be observed in the figure, however. Neither does an ANOVA indicate any significant interaction effect (MS = 40,132; F = 0,897; p = 0,409).

3.2 THE TRIOS SCALE

This scale has 21 tasks or items, and another case of composite summed score is employed. Scores from the initial round of data range between -7 and +21, with a mean of 8,7. The maximum score of 21 thus is reasonable, and negative scores for wrong responses also exist. Figure 4 also shows that a relatively 'normal' distribution of scores.

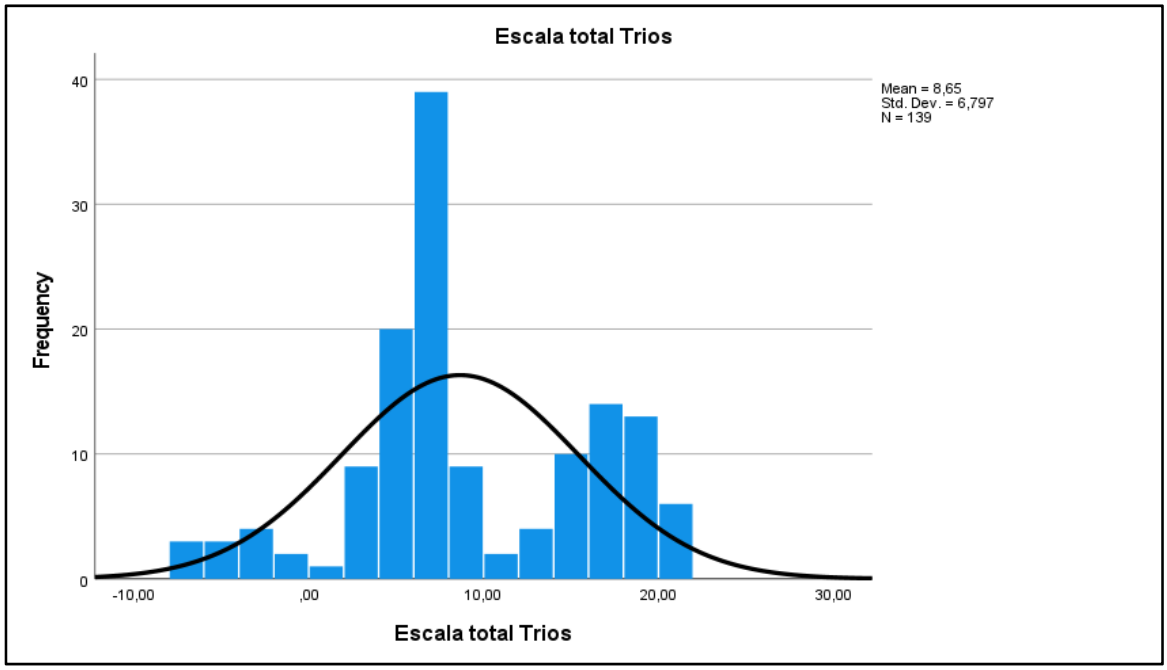

Figure 4: Distribution of the Trios variable at the first time point (pre-intervention)

From the second and third data sets, comparable distributions were found. The respective means were 12,0 and 13,8; with max/min values of -9/+21 and -6/+21.

In figure 5, the means of the control group are consistently higher than those of the intervention group. The differences are small, however, and far from being statistically significant (MS = 139,903; F = 1,533; p = 0,218).

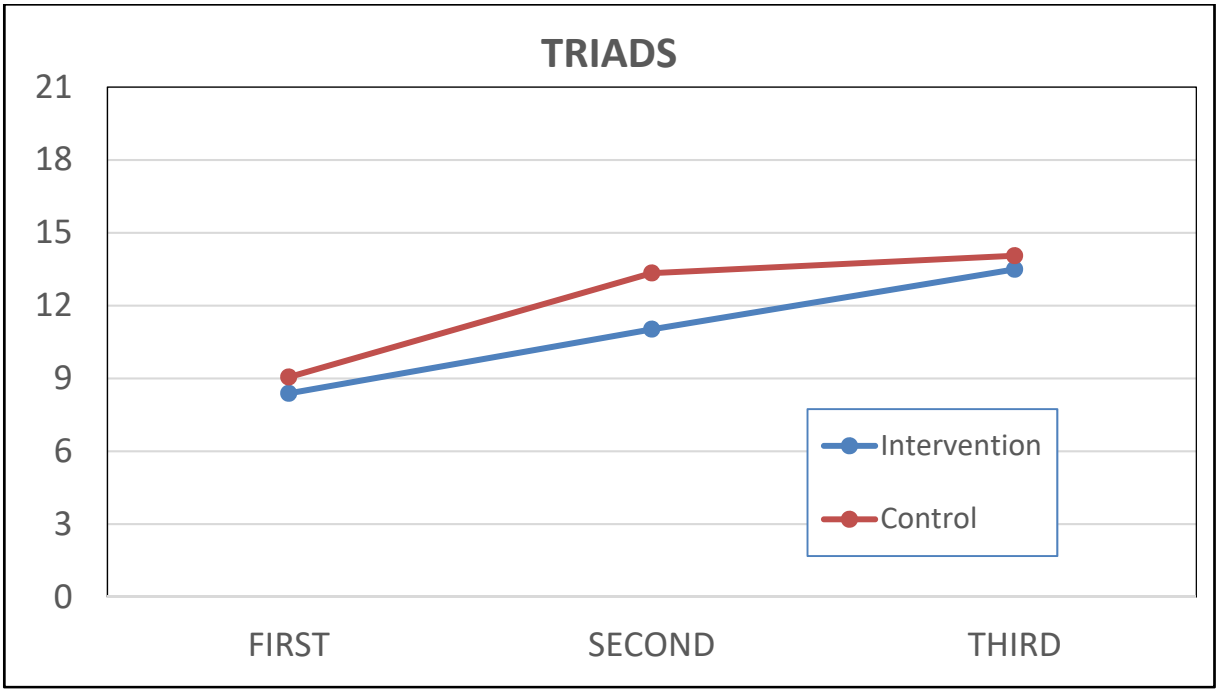

Figure 5: Group means of the Triads variable at three time points

Changes over time do occur also here, however. The means increase throughout the three time points. ANOVA confirms that this change is significant (MS = 893,179; F = 38,993; p < 0,001); and no interaction effect is observed (MS = 32,363; F = 1,413; p = 0,245).

3.3 THE ARROWS SCALE

This scale contains 36 test items, and 'punishment' scores for erroneous responses are employed. The scores for the initial time point thus range between -3 and +35, with a mean of 18,1.

For the second data set, the mean is 22,9; the minimum value 0 and the maximum 35. At the third time point a mean of 23,7 is found, with min. and max. values of +5 and +35.

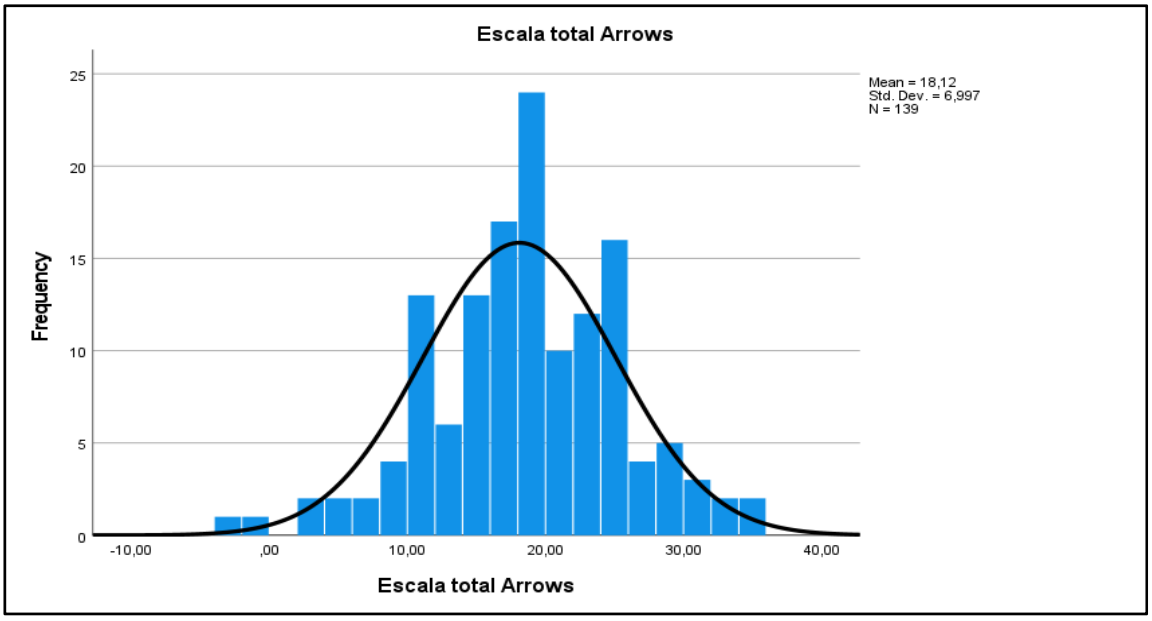

Figure 6: Distribution of the Arrows variable at the first time point (pre-intervention)

Figure 7 suggests that the difference between the intervention and the control group on the Arrows variable is not important. An ANOVA confirms this (MS = 23,843; F = 0,243; p = 0,623). But the means even here do increase with time (MS = 1236,224; F = 53,627; p < 0,001). Finally, there definitely is no interaction effect (MS = 0,315; F = 0,014; p = 0,986).

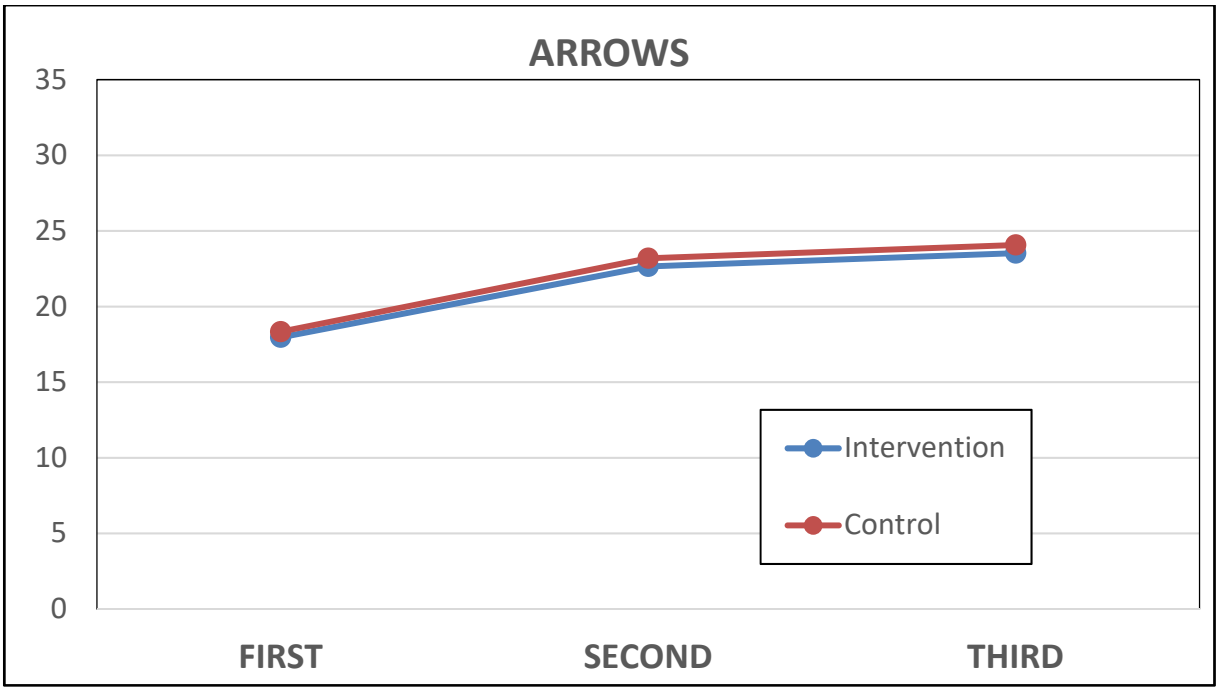

Figure 7: Group means of the Arrows variable at three time points

3.4 THE BINDINGS SCALE

In this scale, there are 27 items or tasks. The mean score for the initial time point is 10. The minimum score is 4, and the maximal score is 24. Since wrong answers here are coded given 0 points (and not -2), negative values do not occur.

In this game, a distinction is also made between *simple* and *difficult* tasks. Correct responses to 'simple' items receive +1 point, while successful encounters with the more difficult are accorded +2 points. It is possible, therefore, to obtain rather high summed scores.

The game also includes a few extremely difficult items, where correct responses would have received +3 points. Responses of this kind are not found in our data sets, however.

In the second round of data, the mean was slightly higher (10,9). The minimum value increased to 5, and the maximum to 30. In the final data set, the mean further rose to 11,1. While the minimum was stable at 5, the maximum score surprisingly decreased to 26.

Firstly, analyses of variance show that the mean increase over time is significant (MS = 38,746; F = 3,751; p = 0,025). Secondly, there is absolutely no difference between the two groups (MS = 0,010; F = 0,000; p = 0,984). The interaction effect also is far from significance (MS = 2,688; F = 0,260; p = 0,771).

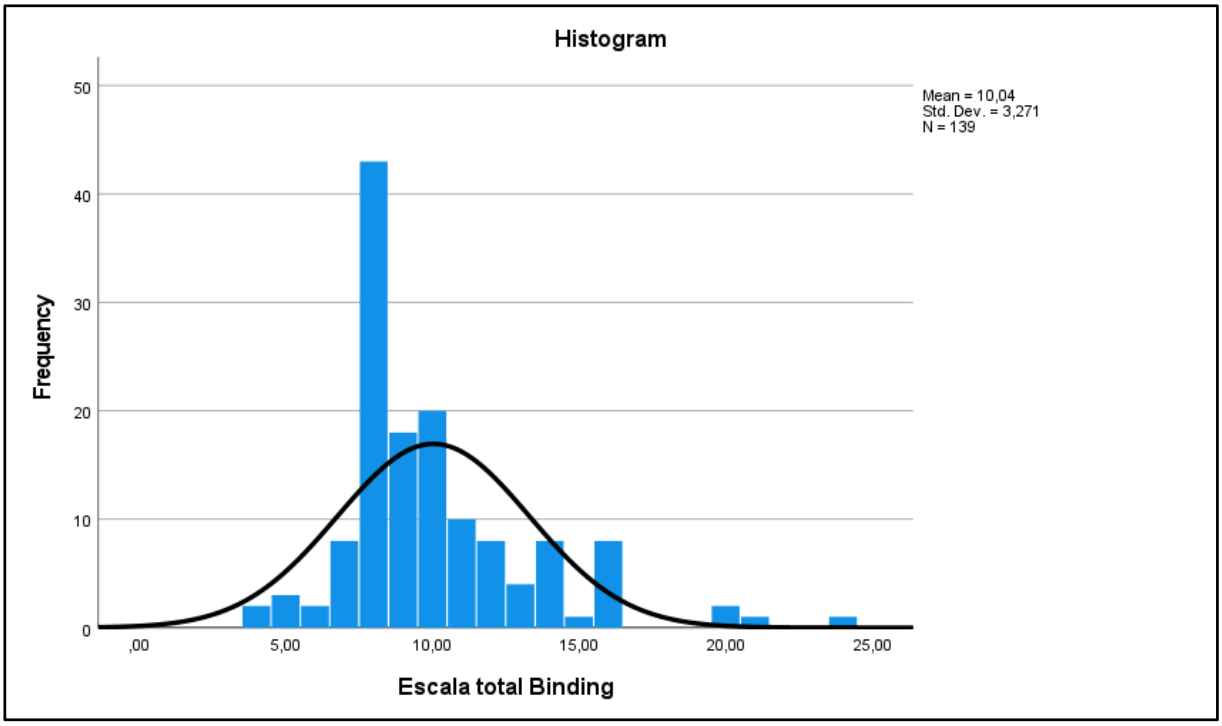

Figure 8: Distribution of the Bindings variable at the first time point (pre-intervention)

Figure 9 confirms the results of the ANOVA. It also shows that the differences between the time points are not really impressive, in spite of the statistical significance of the effect.

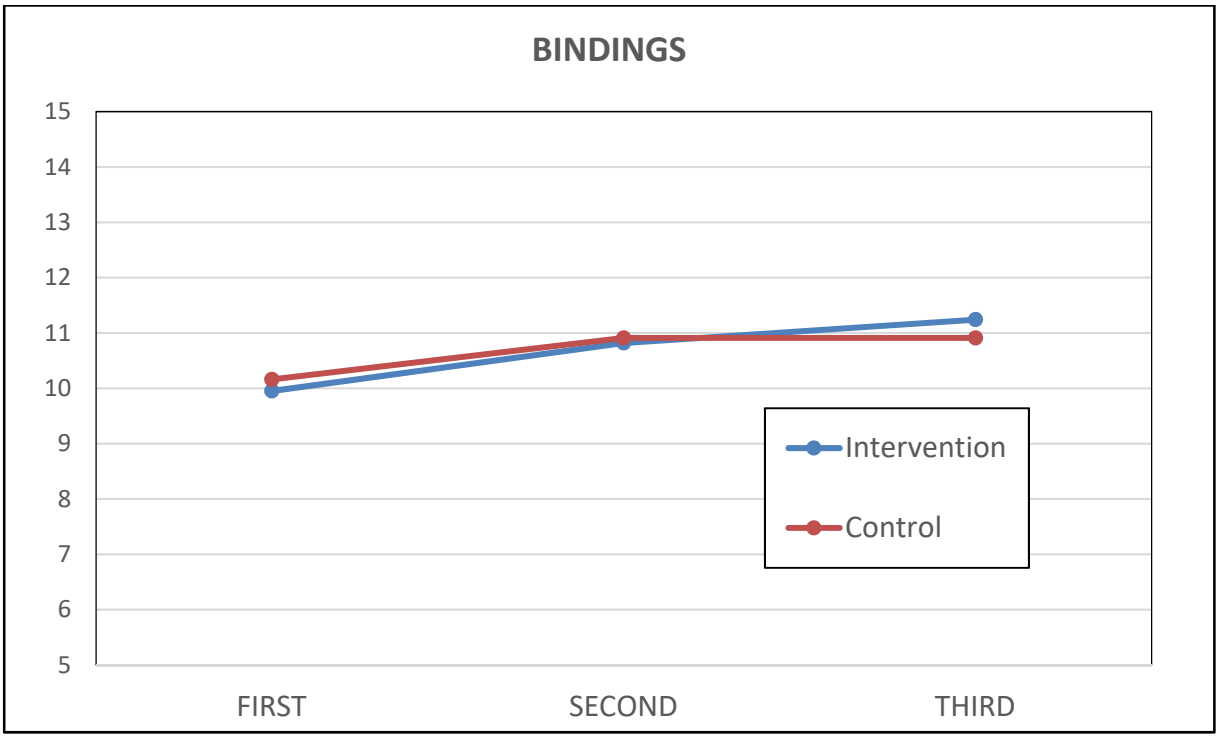

Figure 9: Group means of the Bindings variable at three time points

3.5 COMMENTS TO THE FOUR SCALES

Within the design of the pilot study, then; slightly different results emerge from the four main scales. Table 6 sums up the significant results. The first conclusion is rather clear: Mean scores increase with time. Scores at the second time point are higher than the initial scores, and scores at the final time point are higher than those of the second.

The interaction effect that would have indicated an effect of the intervention, however, does not show up. This is at some variance with the BRIEF results reported in the article by Andersen et al. (2019).

Finally, the control group clearly scores higher than the intervention group on the Cat/Dog-scale (and only there). This slightly odd result appears not to have any obvious explanation, but some badly understood method problem might be a good guess.

Table 6: ANOVA effects (trials x groups), p values

Game/Scale	**Trials**	**Groups**	**Interaction**
Cat/dog	0,001	0,005	0,409
Trios	0,001	0,218	0,245
Arrows	0,001	0,623	0,986
Bindings	0,025	0,984	0,771

3.6 Combined scale (*Suma Z de pruebas*)

Adding together the z-scores of all four measures into one combined score for the entire Yellow/Red test battery is standard procedure for the Yellow/Red test battery. The new score appears not to yield any new information, however.

The distribution of the summed scores distribution looks fairly normal. The mean is 2.54, and the standard deviation is 2.70. And due to missing data, the N unfortunately is only 89.

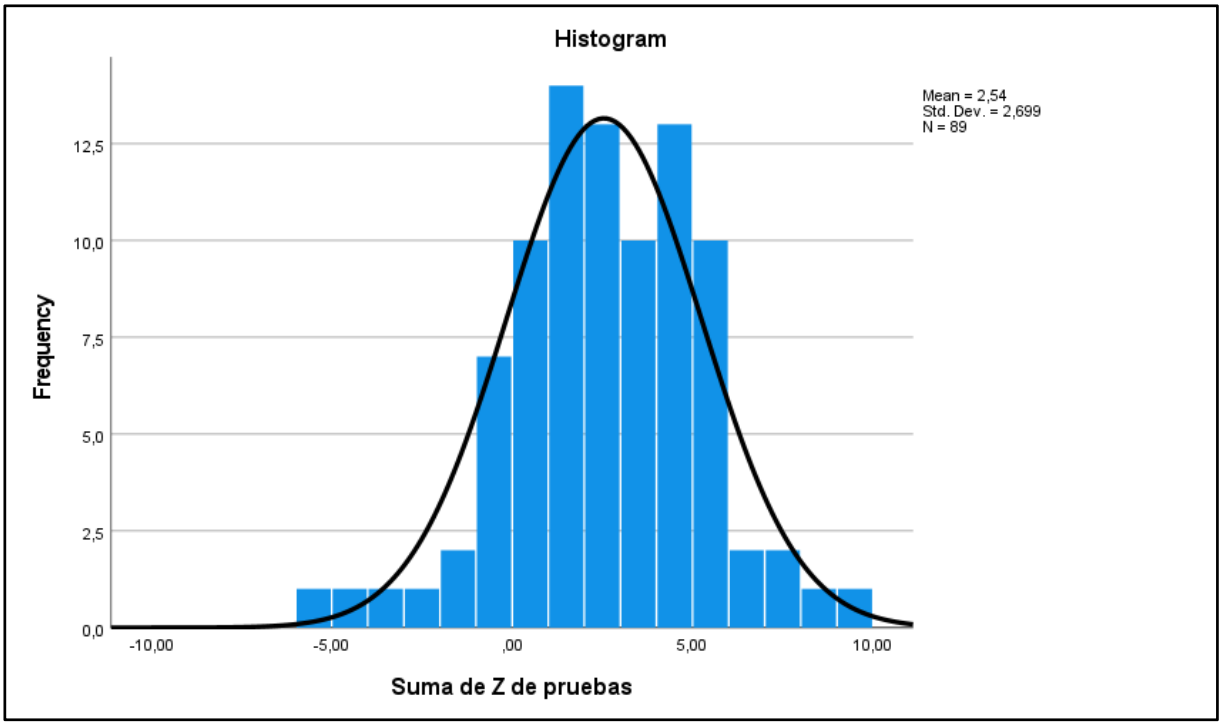

Figure 10: Distribution of combined Yellow/Red score, first data round (pre-intervention)

The changes over time for the new variable is similar to what has been shown for separate scales. As shown in figure 11, its mean increases over time.

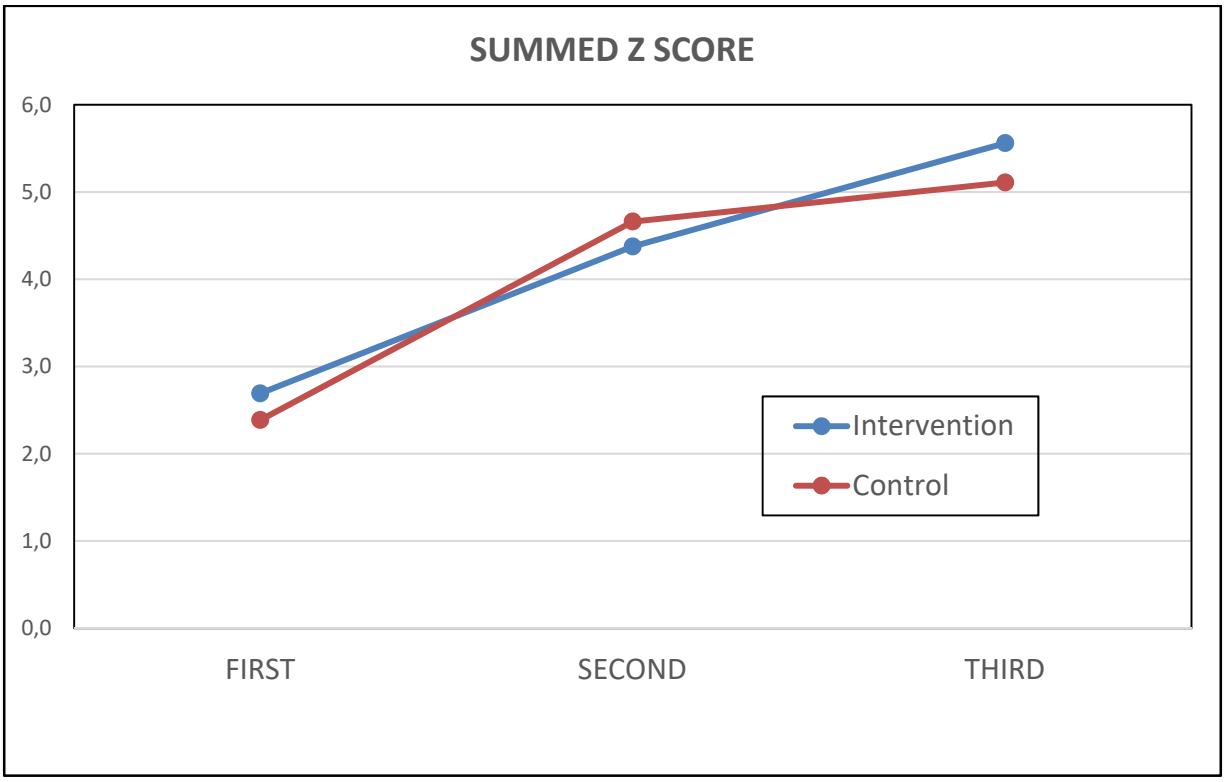

Figure 11: Group means of the combined Yellow/Red variable at three time points

This effect is statistically significant (MS = 183.691; F = 64.138; p < 0.001). There is no significant difference between the intervention and the control groups. The interaction effect between groups and repeats is also not significant.

3.7 THREE-WAY ANOVA: GROUPS BY TRIALS BY MEASURES

A joint analysis, considering the four scales together, may also be of interest. For this purpose, the z-score versions will be used, since they are more easily comparable. A three-way ANOVA was performed, using the three trials and the four game scales as two repeated measurement factors. A third factor is a group factor, i.e., the intervention group vs. the control group.

As one might expect, the results were slightly complicated. But again, there is no general difference between the two groups (MS = 0,409; F = 0,091; p = 0,763). Secondly, the repeat factor of *trials* is significant even here (MS = 45,923; F = 64,138; p < 0,001). Third, there still is no interaction effect between the *groups* and *trials* factors (MS = 0,843; F = 1,177; p = 0,311). Figure 12 on the following page may serve to illustrate these relationships.

So far, then, this combined analysis confirms the results derived from analyzing the four variables separately. The trials factor is the important one, and there still is no interaction effect. There also is no general difference between the two groups; the group difference found when viewing the Cat/Dogs-scale separately does not appear in the combined analysis.

Still, including the four scales as a repeat factor does complicate the findings. The repeat factor of *scales* proves to be a significant main effect (MS = 36,326; F = 32,935; p < 0,001), and there also is a significant interaction effect between the repeat factors of *trials* and *scales* (MS = 3,711; F = 6,568; p < 0,001).

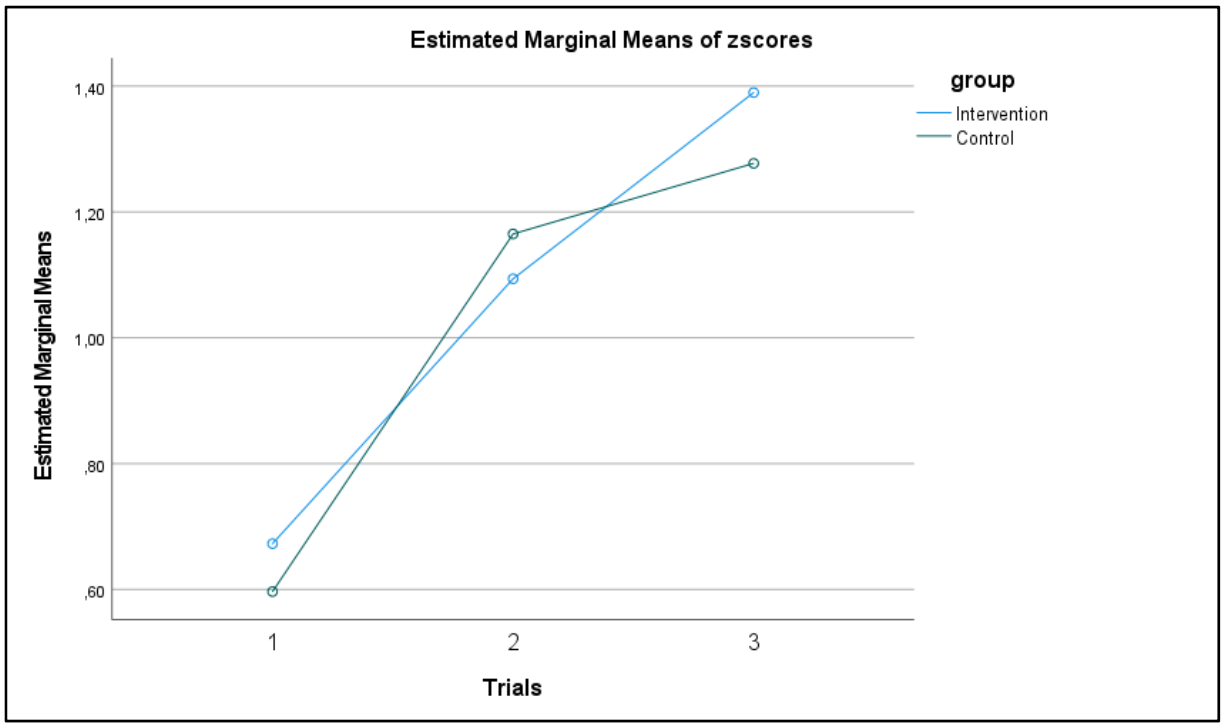

Figure 12: Mean *z* -scores in two groups at three time points, four scales combined

So far, this combined analysis confirms the results derived from analyzing the four variables separately. The trials factor is the important one, and there still is no interaction effect. There also is no general difference between the two groups; the group difference found when viewing the Cat/Dogs-scale separately does not appear in the combined analysis.

Still, including the four scales as a repeat factor does complicate the findings. The *scales* factor proves to be a significant main effect (MS = 36,326; F = 32,935; p < 0,001). There also is a significant interaction effect between the repeat factors of *trials* and *scales* (MS = 3,711; F = 6,568; p < 0,001).

Figure 13 on the next page may serve to illustrate these relationships. But first, please note that the sequence of the scales here comes in an opposite direction. *Bindings* appear as nr. 1 (blue), and *Arrows* as nr. 2 (green). The *Trios* scale is nr. 3 (violet), and Cat/Dog comes last with the number 4 (red).

Clearly, the means of Arrows (green) are higher than the rest, while Cat/Dog (red) is at the bottom. The differences between the four scales are convincing. Although this difference hardly is a problem to the project, a closer look at the scale computations may be in order.

It is also clear that the changes of the four scales over the three time points are partly different. The means of *Arrows* (green) and *Cat/Dog* (red) appear to increase rather evenly with time. *Bindings* (blue), however, tends to increase less than the others, while the means of *Trios* (violet) clearly increase more from the first time point to the second than they do in the next step. Together, these differences make up the interaction effect. The meaning of these observed differences is less clear, however.

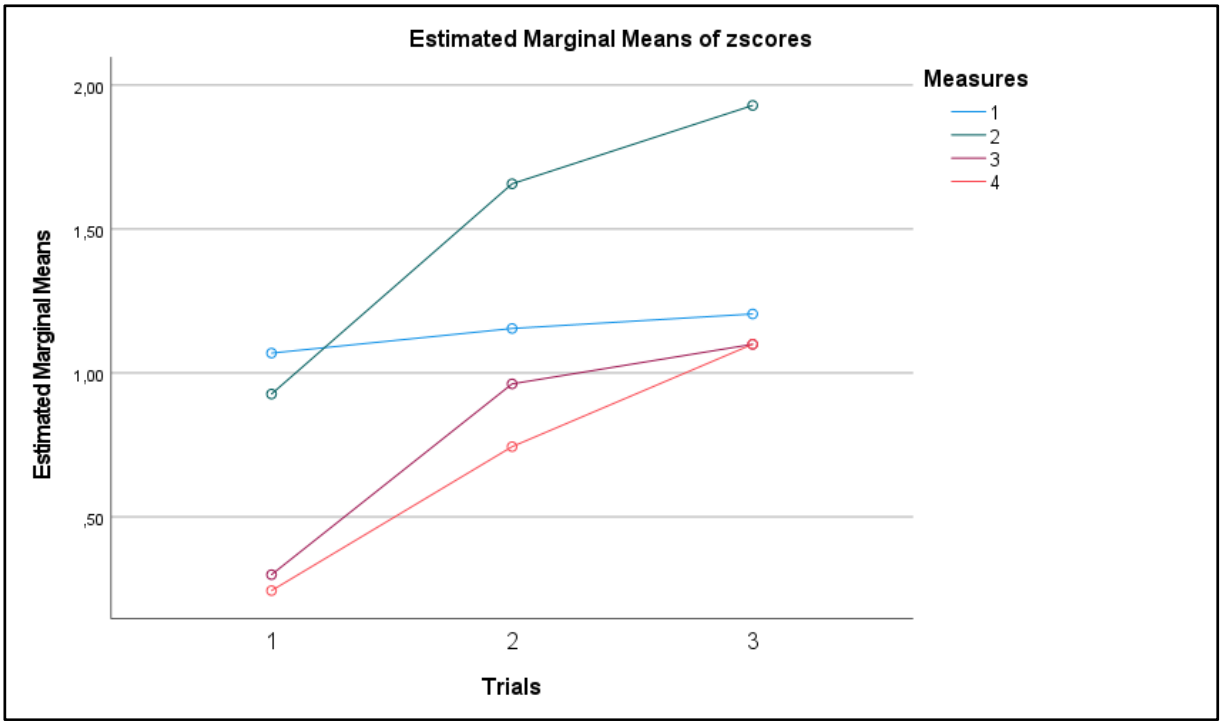

Figure 13: Mean *z* -scores of four scales at three time points, two groups combined

3.8 CHANGES IN RESPONSE TYPES OVER TIME

As shown in Figure 1, different responses to a test item are given different codes or points. In the first three games, e.g., *correct* answers receive +1 points, while *no response* gets 0 and *wrong* responses -2. And in all games, *missing data* means that this data point is removed from all further calculations.

To arrive at a single general score for each game, the points from all items are summed into one common score. This implies, however, that the resulting 'game scale' contains several different kinds of information. The most important perhaps is the 'punishment score' of -2 received for *wrong* responses. This scoring implies a 'double weighting' of the response, compared to the +1 of a *correct* response. It also reduces the summed score more than does the *no response*. Consequently, it serves to disproportionally reduce the general game score. The coding scheme of the fourth game (*Bindings*) is apparently simpler, giving only 0 points to *wrong* responses.

At any rate, the summed score is a *composite*, out together by information about different things. This may raise the question of exactly what constitutes the observed changes over time. Is it simply an increase in the number of *correct* responses, or is it rather a decrease in the number of *wrong* responses? Or do the children grow more cautious with more experience, and abstain from responding when they do are not quite confident? A look at the distribution of response types at the three points in time may improve our comprehension of the increase in the summed scales.

3.8.1 The Cat/Dog scale

Figure 14 shows the percentage of four response types in the Cat/Dog game at the three time points in time. The four response types are mutually exclusive, so that an increase in one response type means reducing others.

The largest changes appear with *correct response* and *no response*. The number of *errors* and *missing data* are relatively small. Also, their reduction over time is more limited than the reduction of the *no response* option. The most interesting may be the increase in *correct responses,* which does not yield a corresponding reduction in the amount of *error* responses. Rather, it is the proportion of *no response* that decreases.

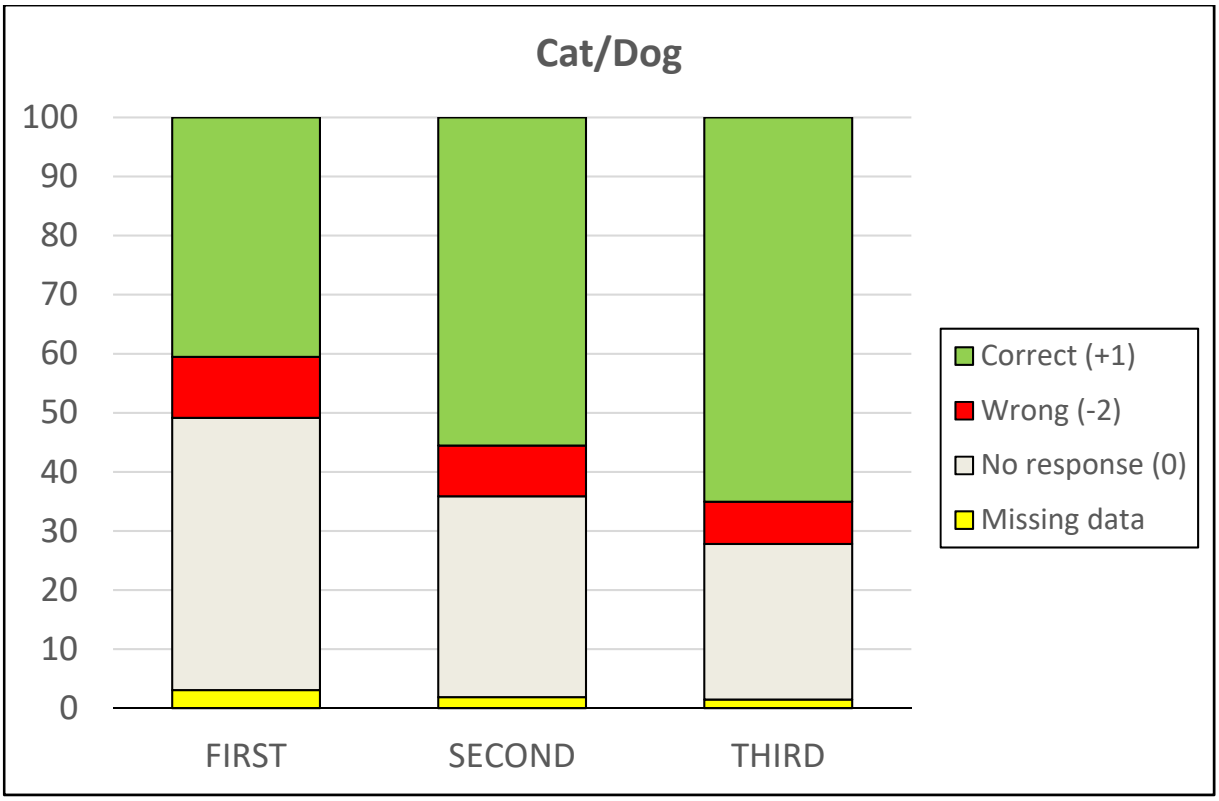

Figure 14: Four response types to Cat/Dog at three time points

3.8.2 The Trios scale

Figure 15 on the next page shows the changes in responses to the Trios game with time. Also here, the most conspicuous change may be the increase of *correct* responses. This increase appears to have gone at the expense of the *no response* and *missing data* responses. And the number of *wrong* responses is small, and it also decreases far less than the percentage of *correct* responses.

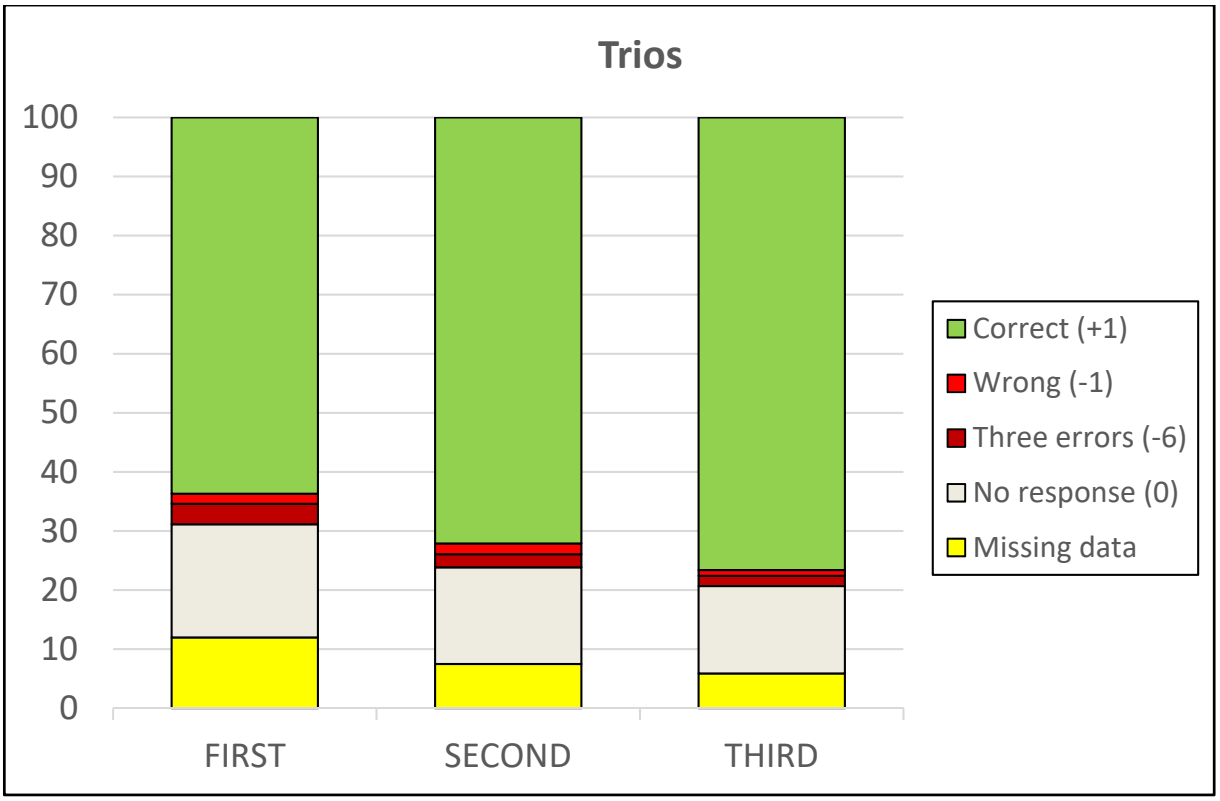

Figure 15: Four response types to Trios at three time points

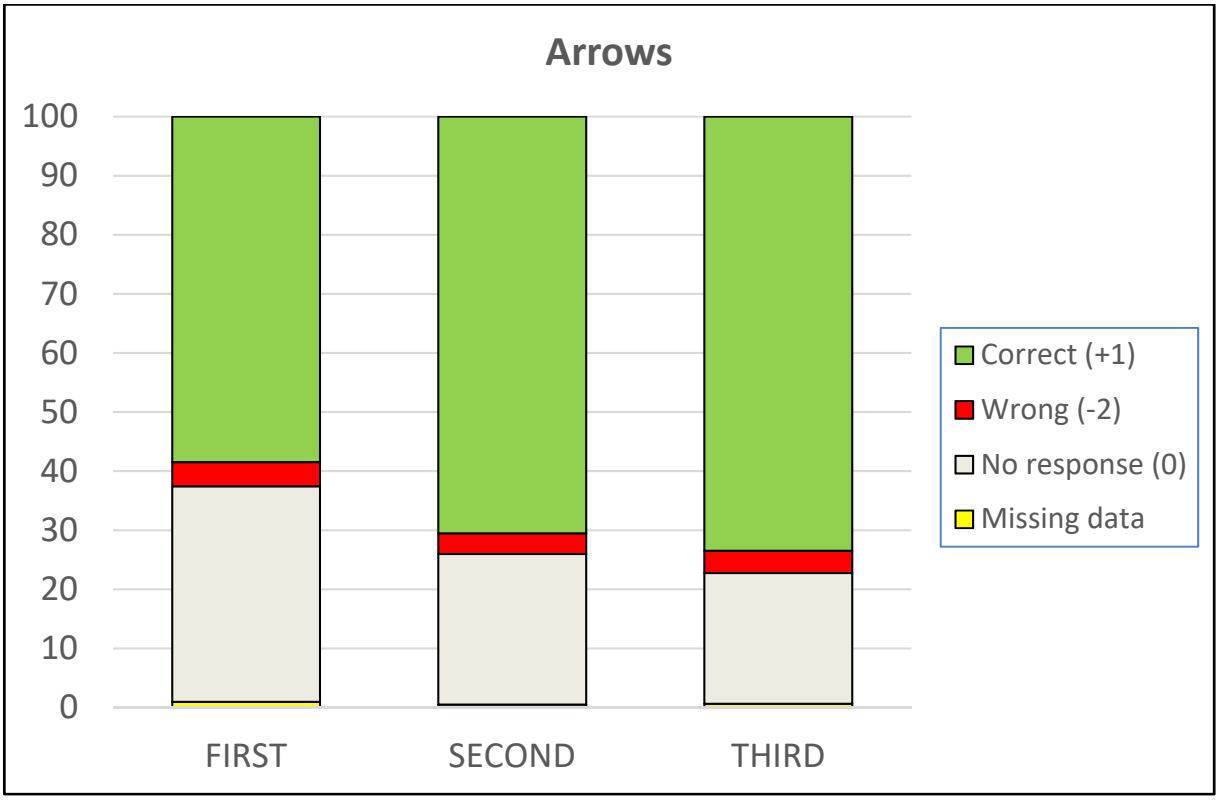

Figure 16: Four response types to Arrows at three time points

3.8.3 The Arrows scale

The increasing number of *correct* responses appears as the central point even in the Arrows game, as shown in figure 16 on the previous page.

Moreover, the increase coincides with a decreasing number of *no response*. The amount of *error* is small, and no important part of the total pattern. It may hardly explain, therefore, the increase of correct responses.

3.8.4 The Bindings scale

The response pattern from this game is rather different from the others. The most obvious discrepancy is the large number of *missing data* in figure 17. This is due to a rule that is specific to this game: the game is terminated after three erroneous responses. The percentage of *missing data* looks rather stable, however, not changing much over the three rounds of data gathering.

It may be more interesting, therefore, to view the increase of *correct* responses over time. Also, this increase appears with the 'easier' (level 1) items, not with level 2. This game also is the only one where an increase in *correct* responses coincides with a decrease in the *wrong* ones.

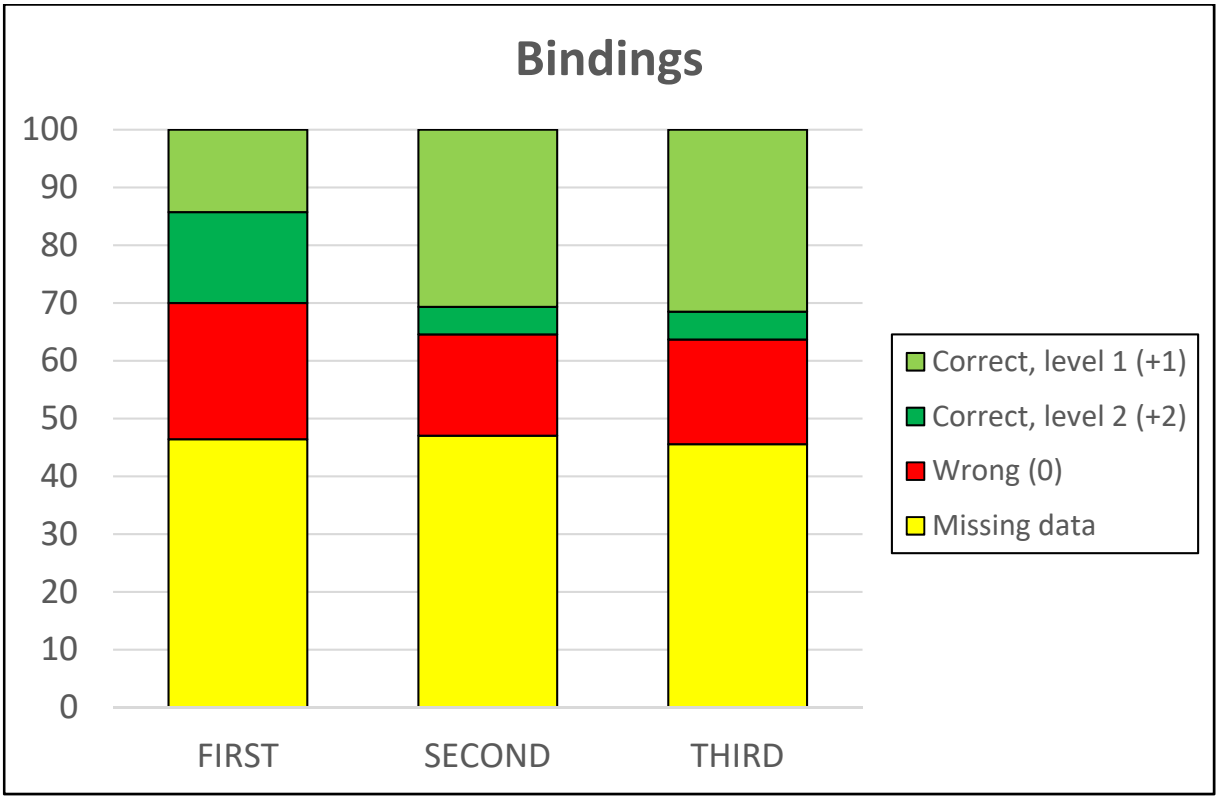

Figure 17: Four response types to Bindings at three time points

3.8.5 Comments on the response patterns in the three data rounds

One main tendency in the total data pattern is that the percentage of *correct* responses increase over time. This generally does not coincide with decreasing numbers of *wrong* responses, however, but with fewer children giving *no response*. In three out of four scales, an increased number of *correct* responses comes at the expense of the *no response* option.

This could challenge our interpretation of the reported increase in the composite summed scores. Does, e.g., increased experience and confidence lead to decreased inhibition – and to fewer cases of *no response*?

At any rate, the improved mean scores over time are hardly caused by reduced numbers of *wrong* responses with negative (-2) scores. The *Bindings* game, however, may be an exception. Only here, decreased numbers of *wrong* responses are likely to play a part in the pattern of changes over time. However, the special procedure for terminating *Bindings* prevents simple comparisons to the three other games.

4. ITEM DIFFERENCES WITHIN THE GAMES

When scores from the single items are added into a combined sum score, it implies that all tasks are given equal weight – whether this is done explicitly or not. If all test items are equally difficult, this choice is fair. It then will not matter much which tasks were mastered, and there is no reason to keep track of the different items.

If one does *not* assume equal difficulty in the items or tasks, however, some more complicated approach may be needed.

4.1 THE CAT/DOG GAME

In this game, the differences between the 33 items may appear as somewhat limited, as shown in figure 18. In this initial round of data, many of the 138 children in the sample had *correct* responses to many items. Many also gave *no response*. The number of *wrong* responses clearly is smaller, and very few data points are *missing*.

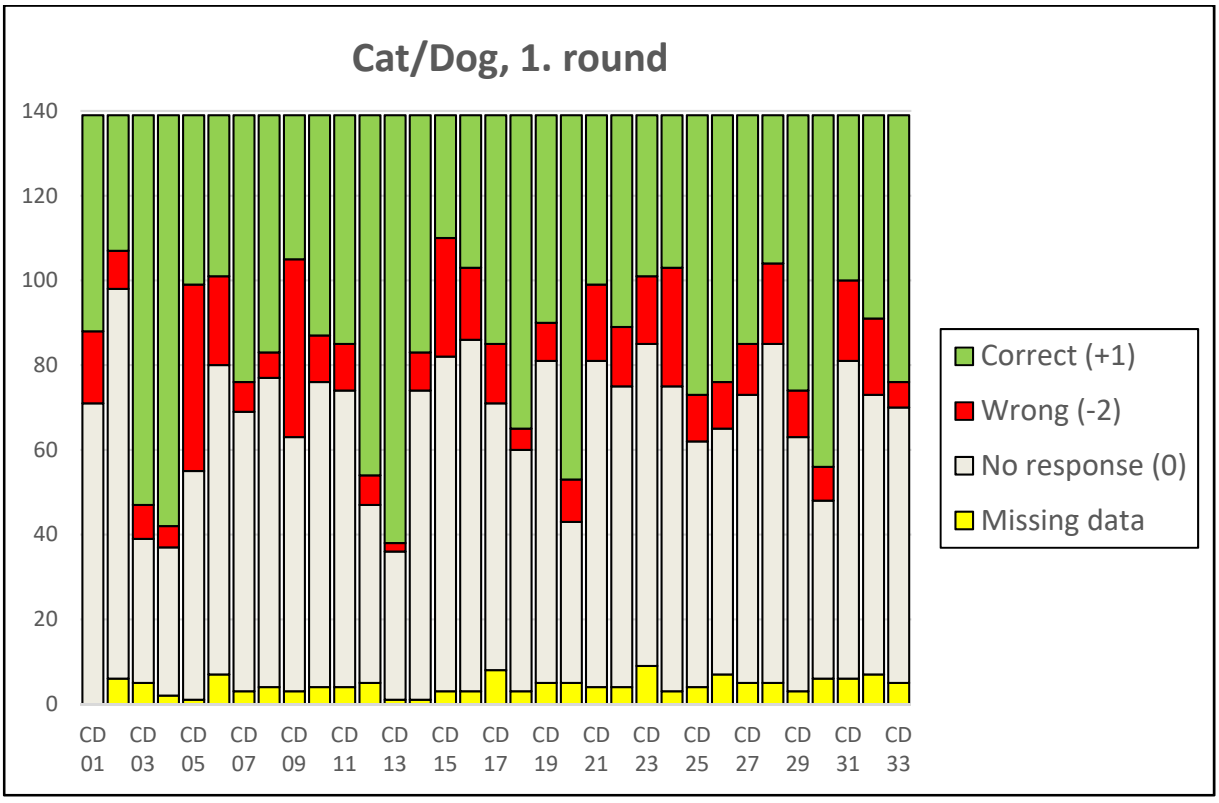

Figure 18: Response pattern to 33 items in the Cat/Dog game, first data round

However limited, the different difficulty of the 33 tasks may also be observed. Items 1 and 2, e.g., consistently receive fewer *correct* responses than do items 3 and 4. This applies to all three time

points. Also, item 4 is always 'easier' than item 3, with a higher number of *correct* responses. This type of difference is easily found in figures 16 -18, suggesting that some attention should be paid to consistent inequalities among the items or tasks. This general response pattern does not change much in the two subsequent data sets, as shown in the two next figures (19 and 20).

The trends of figure 13 should also be kept in mind, however. There, a combination of increasingly *correct* responses and diminishing numbers of *no response* was more evident, while the number of *wrong responses* did not increase. When items are viewed separately, this is less apparent.

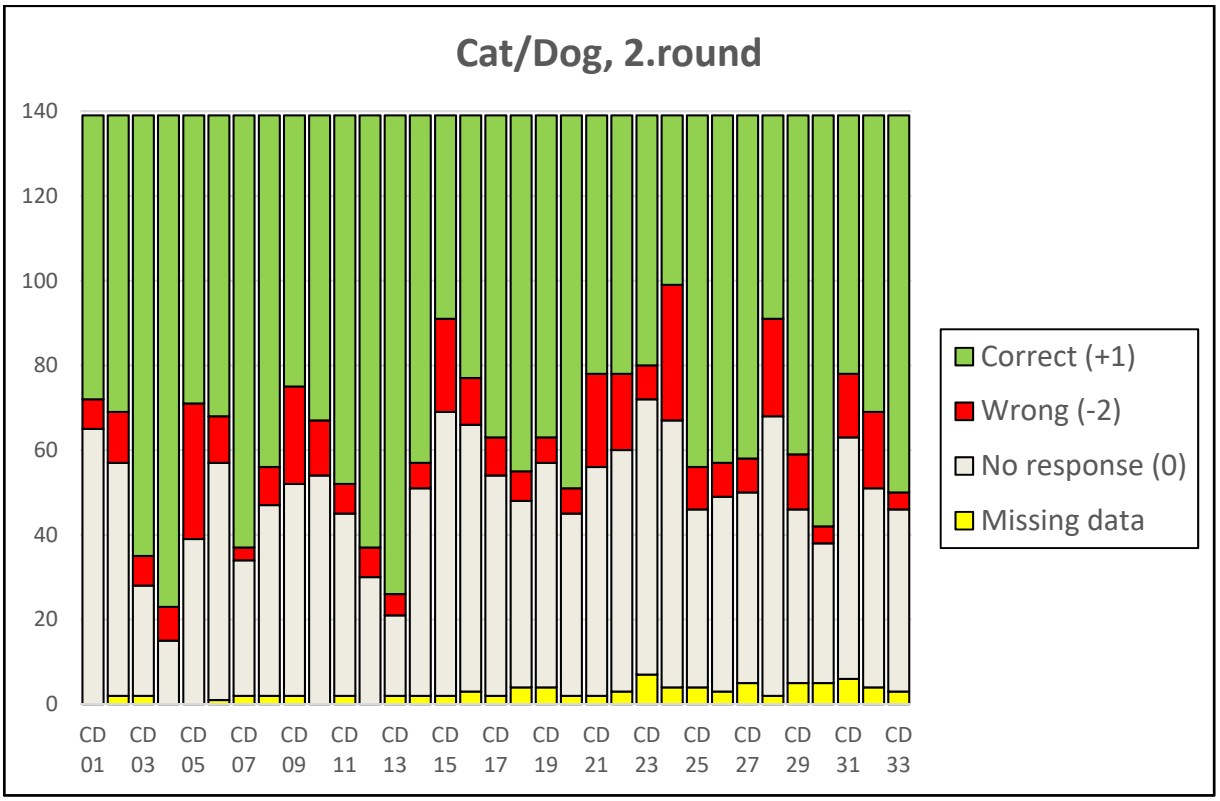

Figure 19: Response pattern to 33 items in the Cat/Dog game, second data round

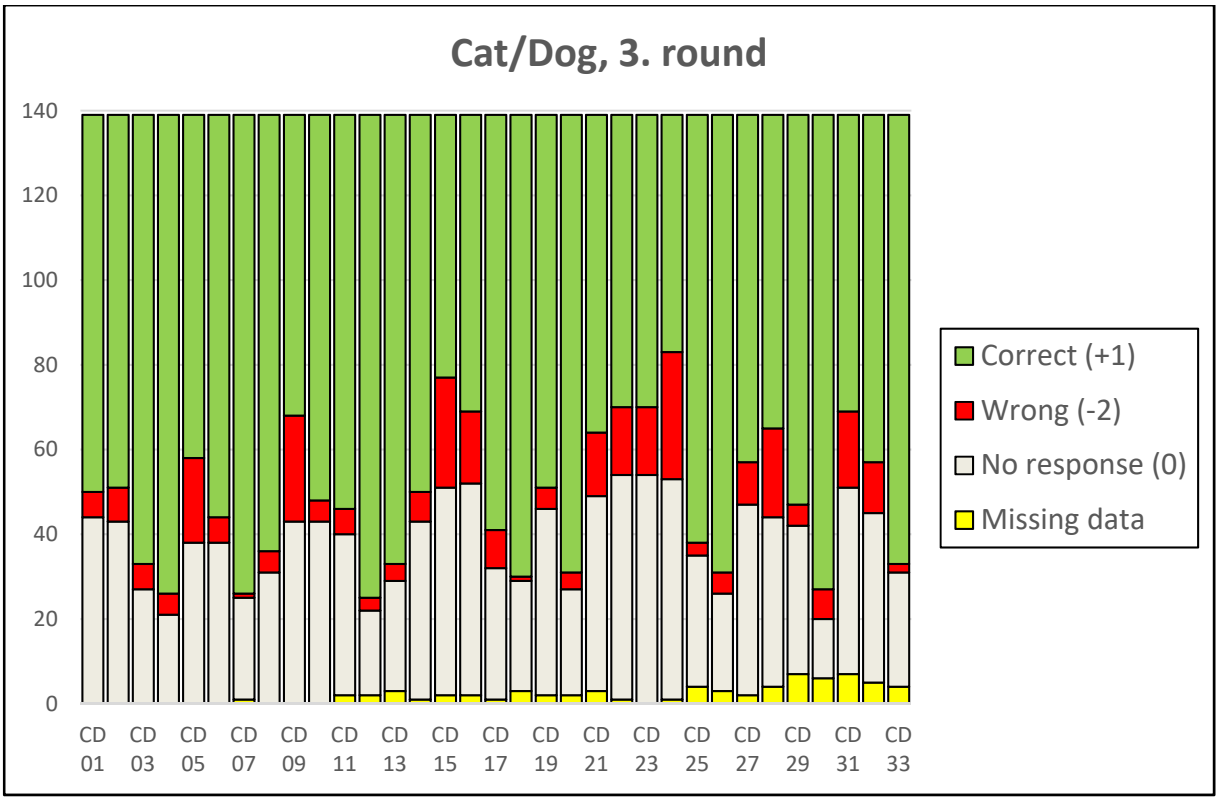

Figure 20: Response pattern to 33 items in the Cat/Dog game, third data round

4.2 THE TRIOS GAME

Also in this game, clear and consistent differences of difficulty between the items appear, as shown in figures 21 through 23. Some tasks are easy, yielding many *correct* responses. Others are more difficult and give higher numbers of *no response* or *missing data*. Simple *wrong* responses are also rare in this game.

Several types of interesting item differences may be observed, however. Item 11 clearly stands out, by giving an unexpectedly large number of responses with *three errors*. Besides, items 12 through 15 receive disproportionally high numbers of *missing data*. And for some reason, items 16 through 21 only produce *no response* and *correct* responses. Finally, items 1 through 8 are demonstrably the easier items of the set, yielding uncommonly high rates of *correct* responses. These differences are easily observed though all three data rounds and leave no doubt about the existence of unequal item difficulty in this game.

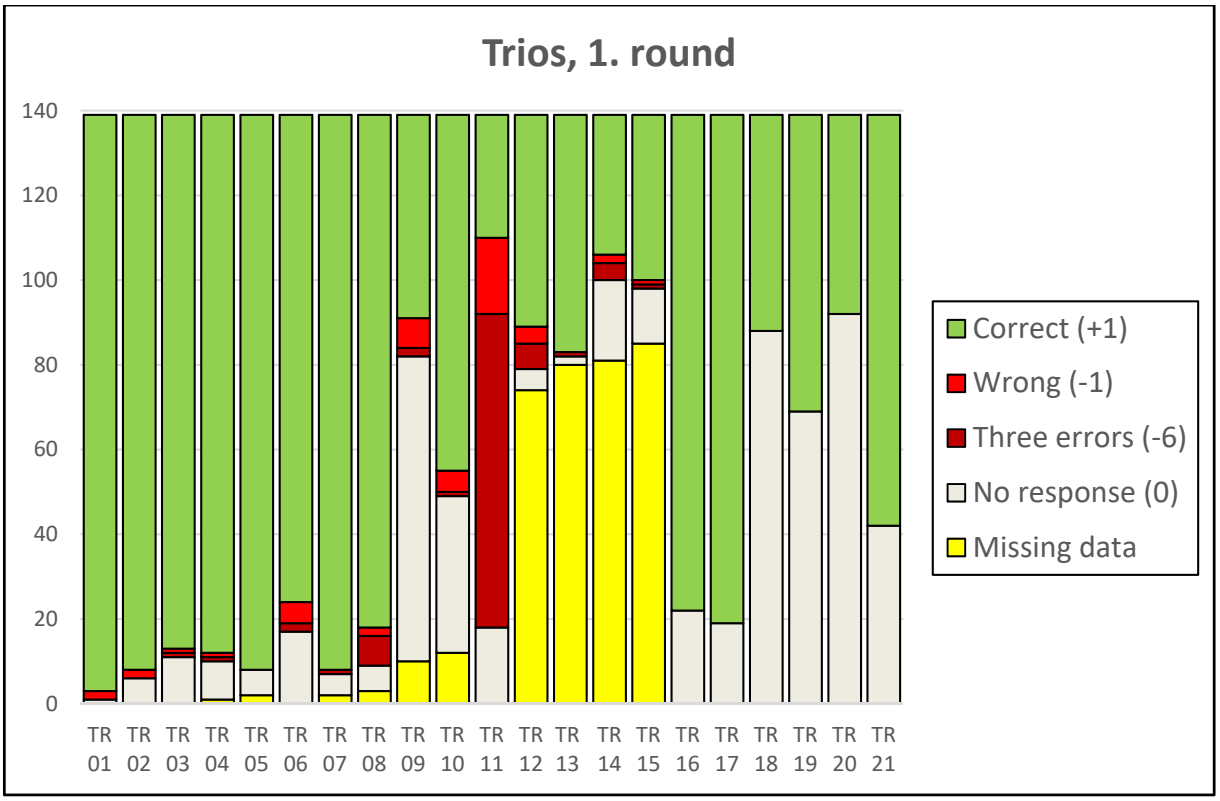

Figure 21: Response pattern to 21 items in the Trios game, first data round

The three response patterns should be compared to the change in mean scores that was displayed in figure 13. It is true that the share of *correct* responses does increase from the initial through the last round of data, coinciding with increasing mean scores. And yet, the general score differences also serve to conceal interesting item differences. With item differences of this magnitude, there is unexplained variance in the material that may warrant closer scrutiny. Important differences and relations may exist, that summed scores and their means do not explain.

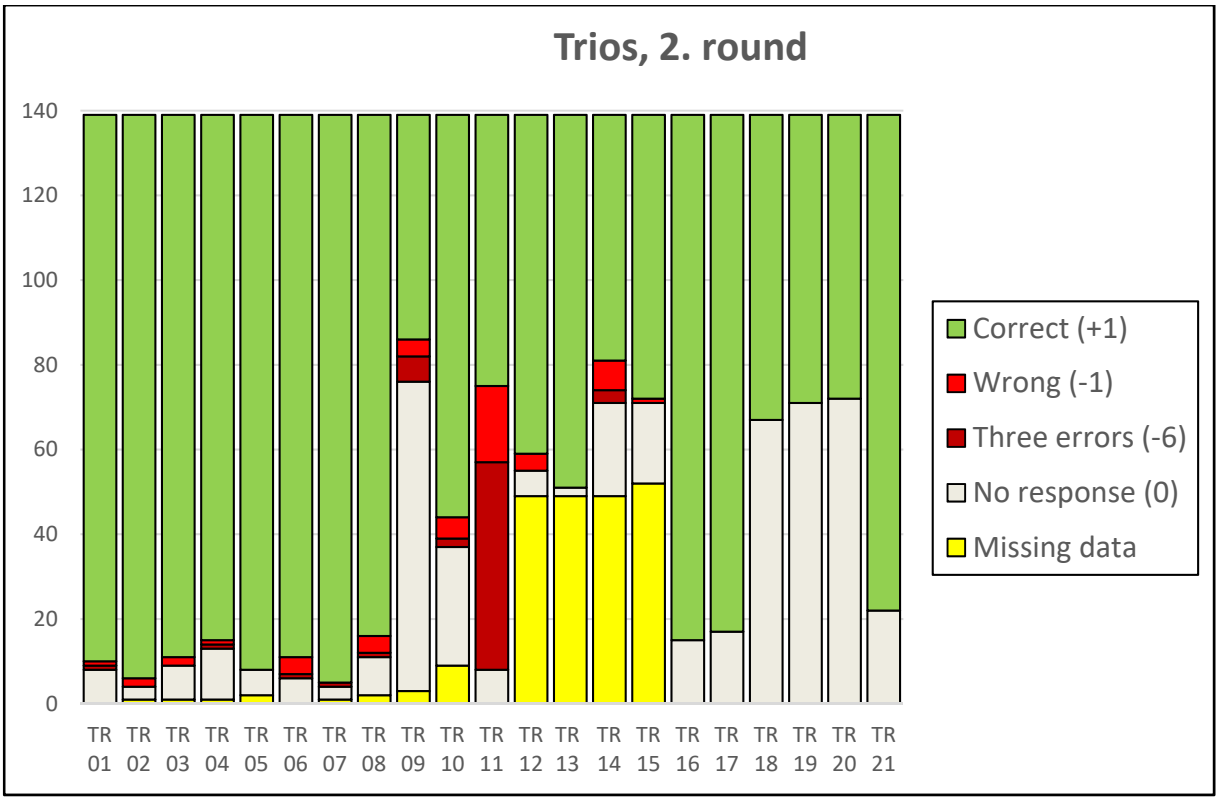

Figure 22: Response pattern to 21 items in the Trios game, second data round

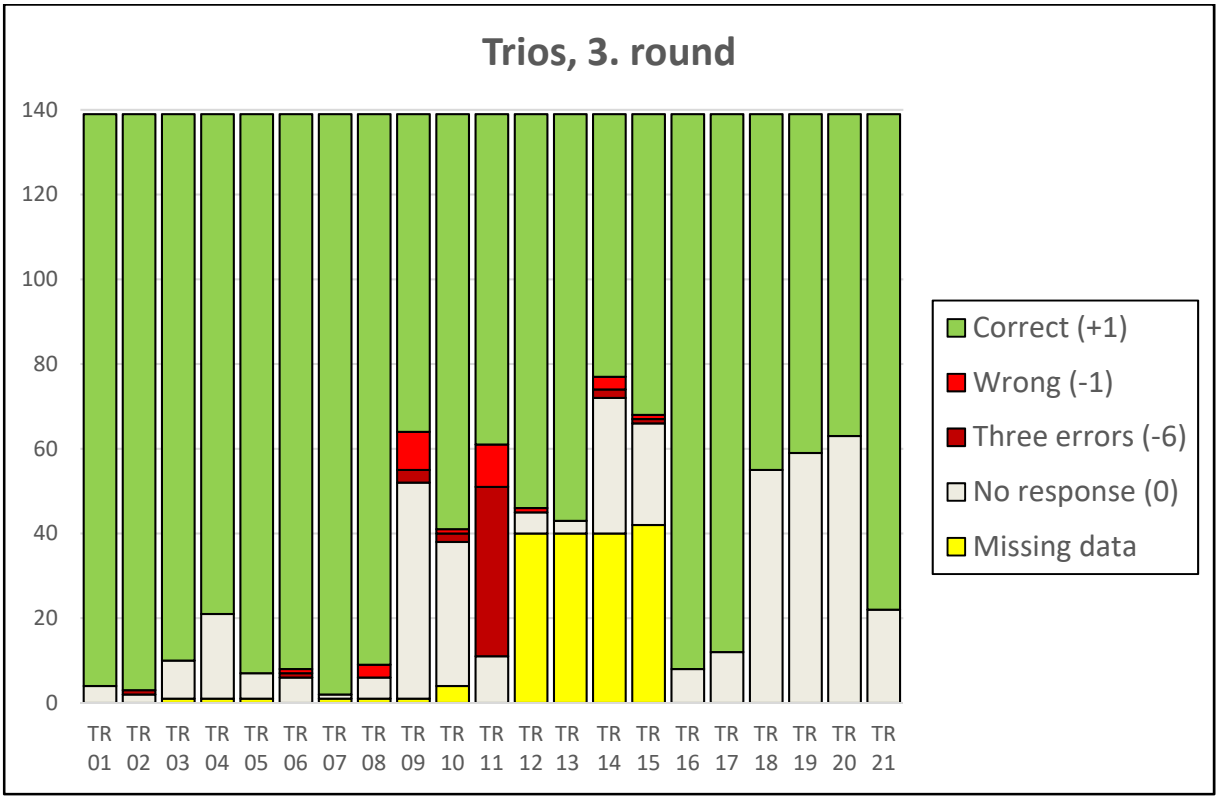

Figure 23: Response pattern to 21 items in the Trios game, third data round

4.3 THE ARROWS GAME

In the Arrows game, the response pattern has much in common with that of the Cat/Dog game. As indicated by figure 24, the 36 Arrows items are not equally difficult. The initial 16 tasks, e.g., appear to be easier than the following, while the last half of the game produces a relatively high number of *no response*. *Wrong* responses are quite rare. But, perhaps paradoxically, they are mainly found on tasks with a relatively high number of correct responses.

Besides, a 'stairways' pattern is evident in the final half of the game, reminiscent of the stepwise structure of Guttman scales (Nunnally & Bernstein, 1994). Item 17, e.g., rarely yield *correct* responses, but the number of *correct* ones gradually increase from tasks 18 through 21. Similar patterns also emerge with the tasks 23 through 26 as well as 27 through 29.

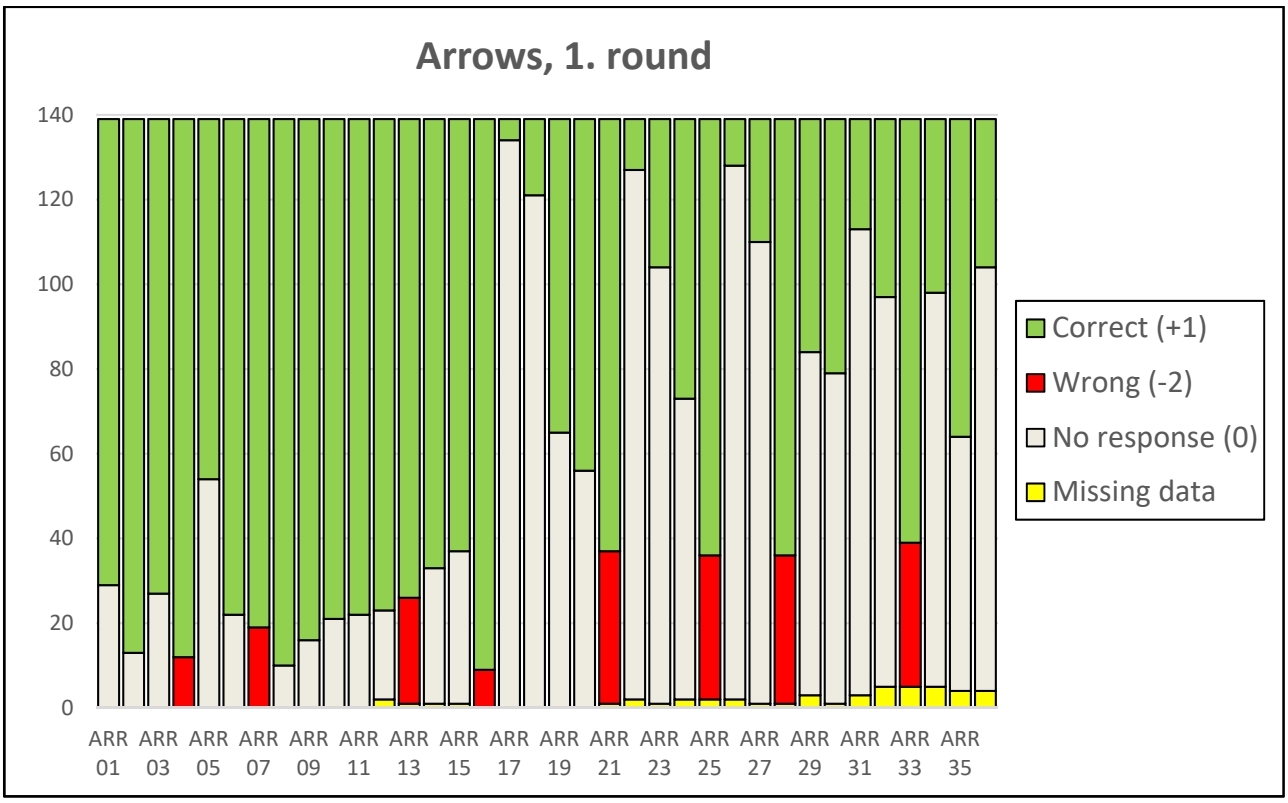

Figure 24: Response pattern to 36 items in the Arrows game, first data round

Again, the pattern is repeated in the next two waves of data. The initial items receive the most *correct* responses, and *no response* also is rather common. *Wrong* responses are rather infrequent, and the stepwise pattern is easily recognized in figures 25 and 26.

At the same time, one may observe that the share of *correct* responses increases a bit from the first to the second data round, and then further into the third. This matches well with the changes that were indicated in figure 14.

The large differences between the items raise the same questions about unexplained variance as did the RESPONSE patterns of the Dog/Cat and Trios games. What is really constitutes the nature of the item differences, and how do the differences influence the general response patterns?

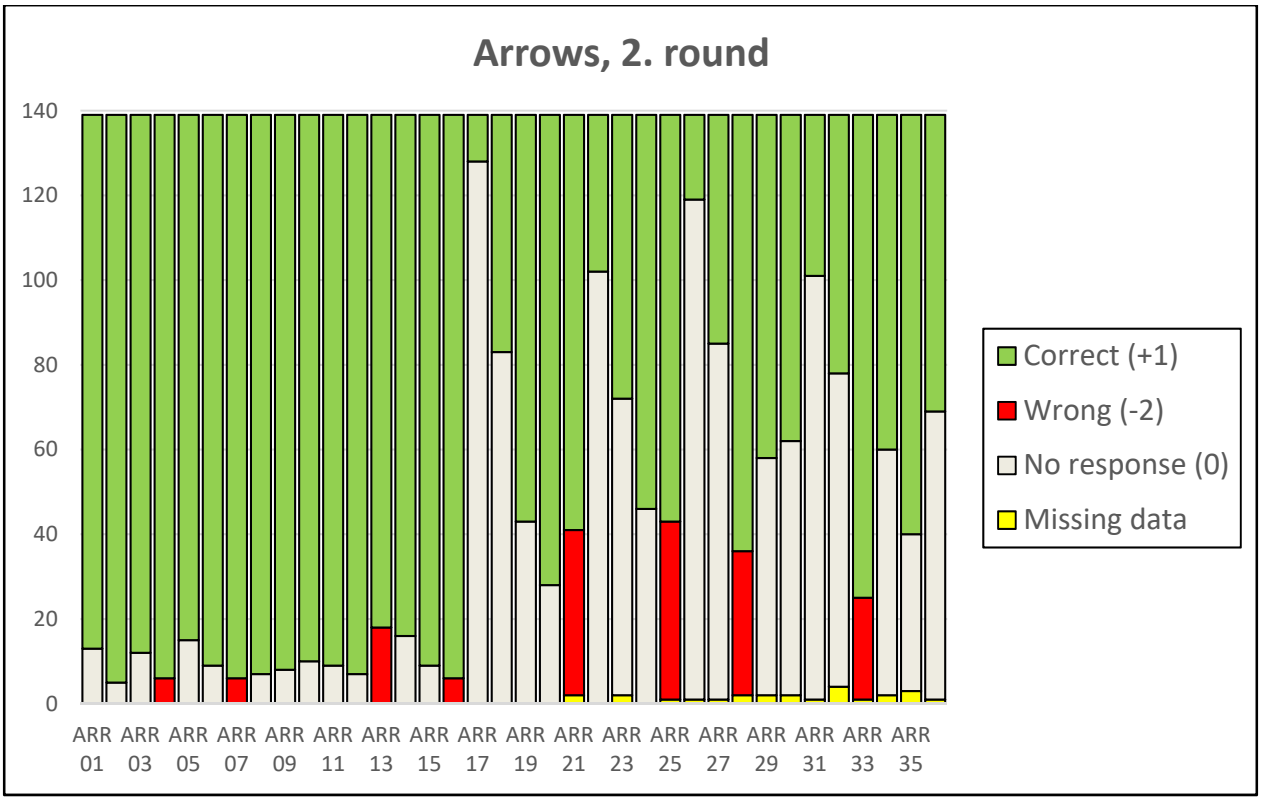

Figure 25: Response pattern to 36 items in the Arrows game, second data round

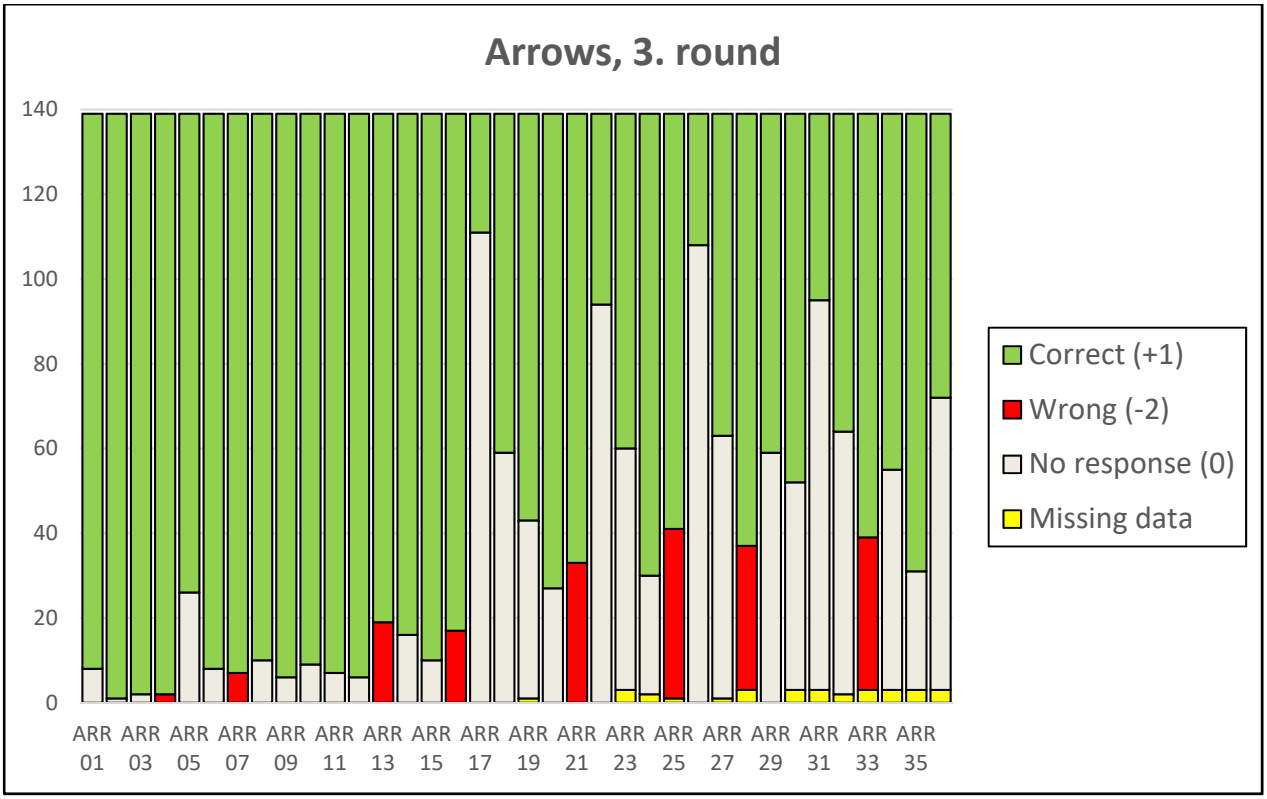

Figure 26: Response pattern to 36 items in the Arrows game, third data round

4.4 THE BINDINGS GAME

The response pattern this final game contains an exceptional number of lacking responses (*missing data*). There also are more *wrong* responses than given in the previous games. The two grades of '*correct*' responses should also be noted, yielding different number of points. The test really employs three levels of difficulty, but no items at the third (high) level were solved in our sample.

At any rate, large item differences are evident in figure 27. Consequently, the distribution of response types greatly varies between the 27 items. The difficulty of the items obviously run all the way from easy (*correct* response from most respondents) to difficult (no *correct* response).

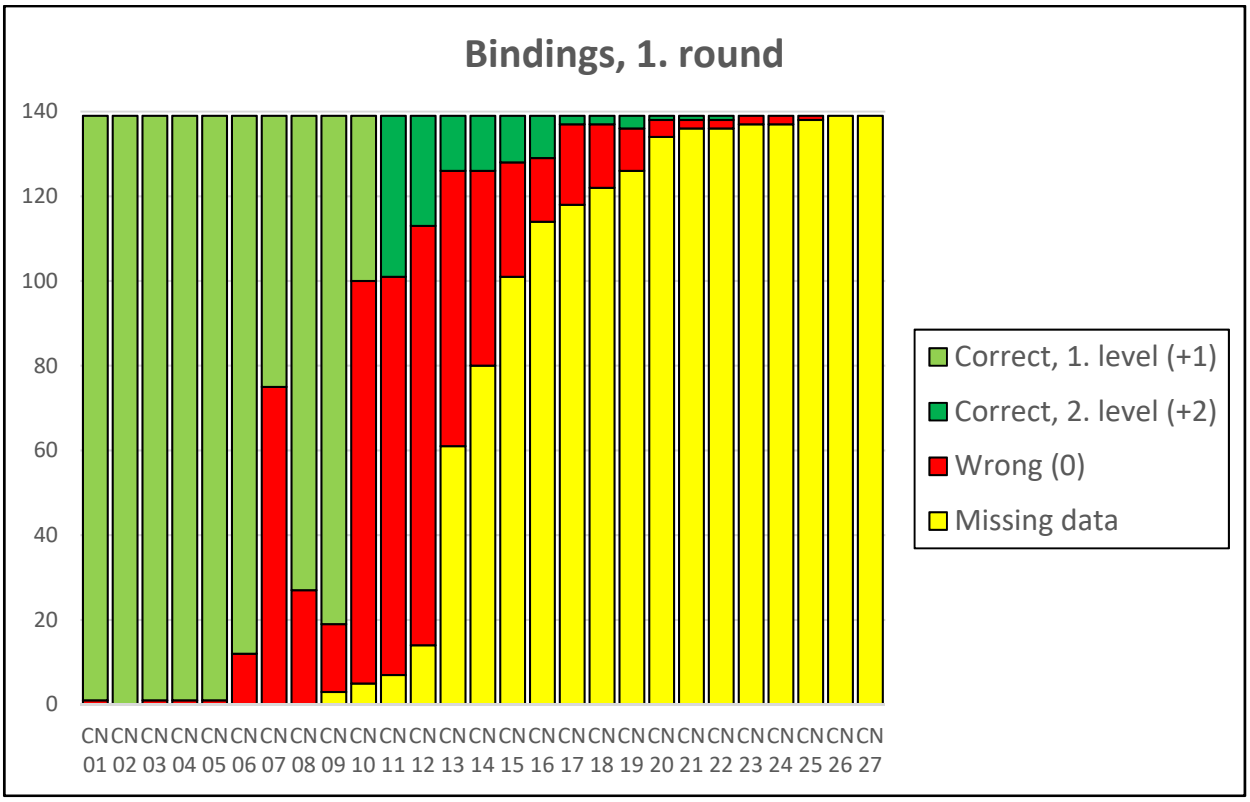

Figure 27: Response pattern to 27 items in the Bindings game, first data round

The most striking feature of this distribution, then, is the different difficulty levels of the 27 items. We lose information also in this game, when the scores from all items are lumped into one common sum. And the issue of 'suppressed' variance is no less relevant than with the previous games; what could the differences between tasks tell us? Could information about the item differences help improving our understanding of the development of children's executive functions?

And even here, comparisons to the general trends (figure 17) are relevant. It may be less easily discovered in figures 28 – 29, but they also show that the number of *correct* responses increase over time. It may also be possible to discern a slight decrease in the share of wrong responses.

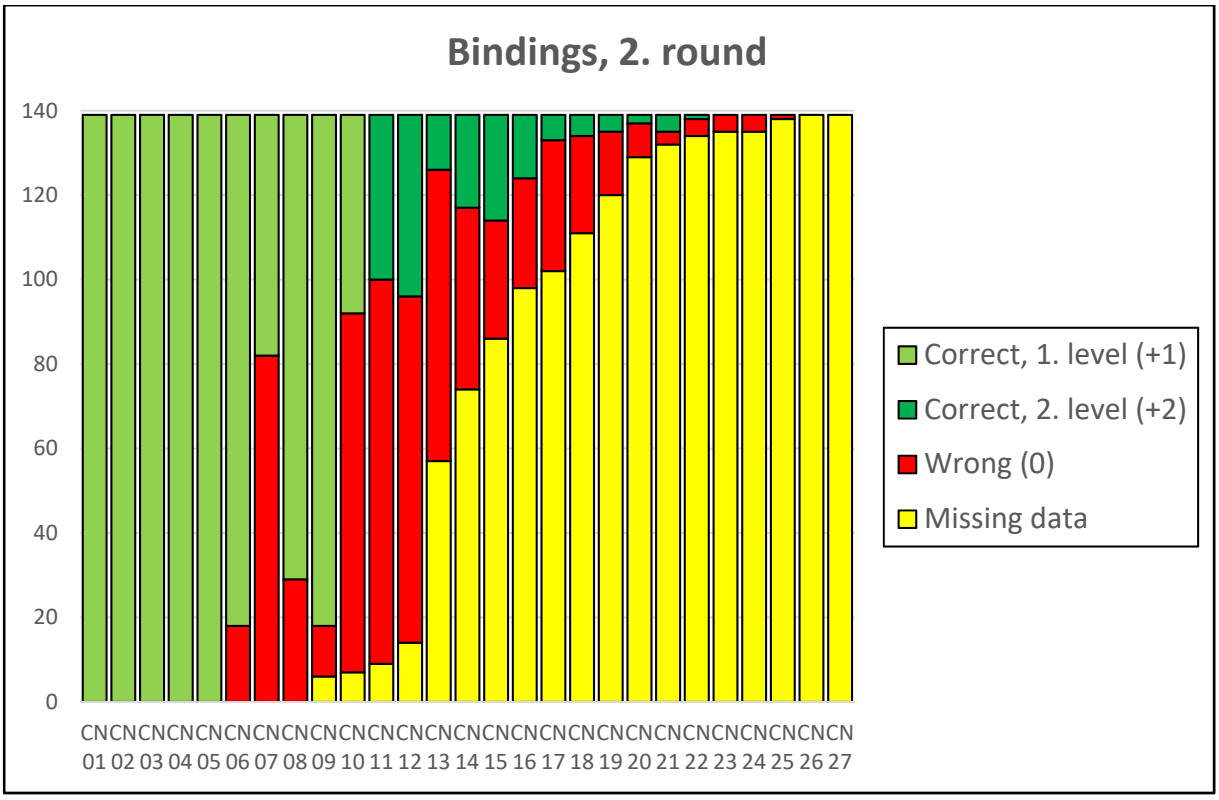

Figure 28: Response pattern to 27 items in the Bindings game, second data round

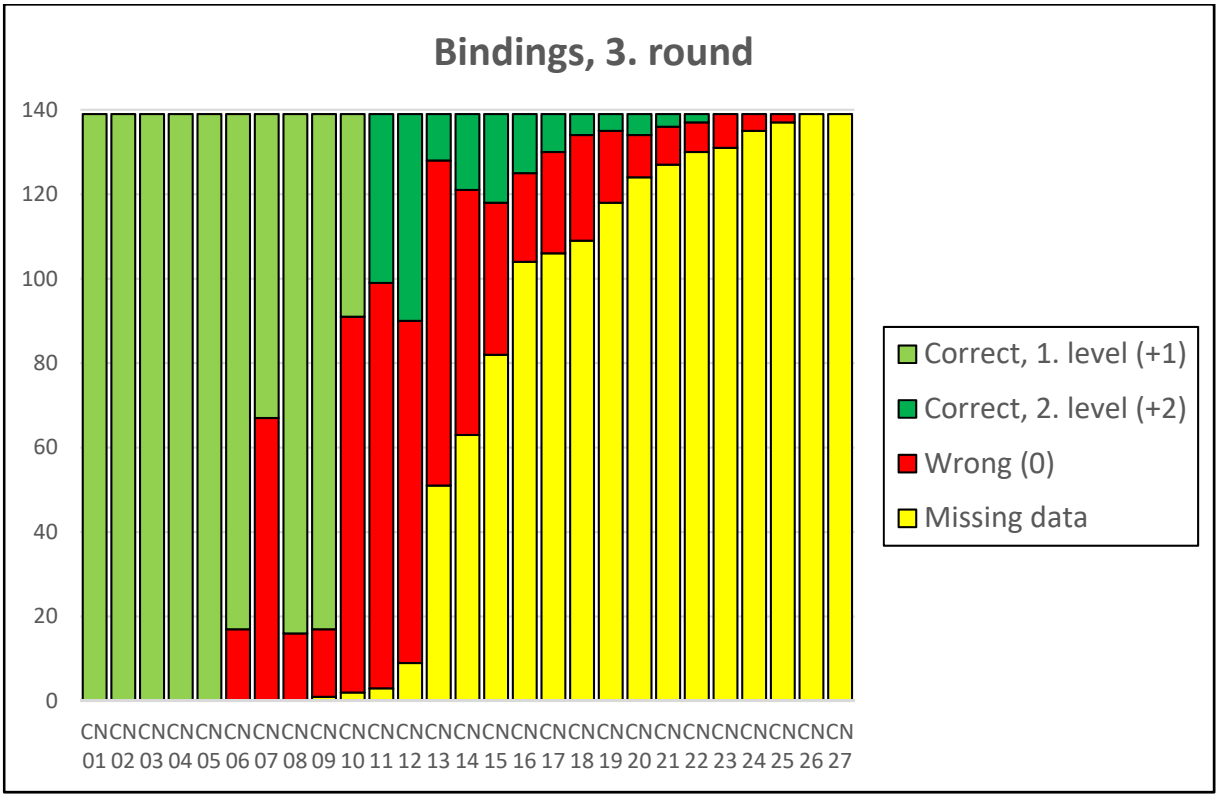

Figure 29: Response pattern to 27 items in the Bindings game, third data round

5. SOME PRACTICAL OPTIONS

As we have seen, the measurements of the four Yellow-Red scales may raise a few questions. Alternative approaches to methods thus may be of interest. Improved reliability and validity of data from the four games would of course be desirable, even if the path to such improvements is not immediately clear.

It a hardly advisable, however, to alter the basic procedures of the games – or the sequence of their tasks. Substantial parts of these algorithms have been transformed into automated tablet programs, so that changes are likely to prove costly. Besides, a lot of data have been gathered with the present set-up, and compatibility with future data sets is a major concern. This prevents major changes in the present game formats.

It may be simpler, however, to have a close look at the *coding* of the responses. Here, some experimentation with alternatives may be easier. Adjustments of the coding algorithms need not involve major changes in the game procedures. It may be useful, therefore, have a look at the pros and cons of alternative coding or classification schemes for the responses in the games.

As shown in figure 1, responses to the items are coded in partially different ways in the four games. For the sake of simplicity, however, the present discussion will be limited to the Dog/Cat game. But certain points and questions are of more general relevance. Therefore, they apply to all games and their scales, but in partly different ways.

As shown several times, the scoring of all the games collects *information about partly different things* into one common, summed score. In the Cat/Dog game, the algorithm first classifies the response to each item into *correct* (+1 points), *wrong* (-2), *no response* (0) or 'premature'. All premature responses (right or wrong) are immediately recoded into a *no response*, also receiving 0 points. This excludes the possibility, e.g., of scoring *both* correct and premature. The scoring thus assumes *mutually exclusive* codes for *correct*, *wrong* and *no response.* It also assumes that *wrong* responses are twice as important (-2) as *correct* ones (+1). Nevertheless, the points from all 33 codes finally are simply summed into one common score for the entire game.

For each game, **the number of correct responses** may be viewed as correct and relevant information about the child's level of coping or mastery. Many such responses show that the tasks generally are solved in a satisfactory manner, contributing to a high total score for the game. In contrast, a smaller rate of correct responses means that several tasks have been too difficult. The smaller number of *correct* responses thus contribute less towards a clearly positive score.

This relationship between the measurement (score) and the task performance (item) is simple and direct. This may be viewed as an argument for keeping the information about *correct* responses separate from data on other responses. The other response types, however, may still be relevant and interesting for other purposes.

Wrong responses («punishment» points of -2), e.g., also depend on the task being solved or not, and it hardly is unreasonable to let the negative figure reduce the summed score. The *wrong* score of -2, however, implies twice the weight of the +1 for the *correct* response. The «punishment» points thus have a double effect. This implicit weighting is not easy to explain: Exactly how does one arrive at -2, and not -1 or -1,5?

The **no response** code may also hide interesting responses**.** The lacking responses also contribute into the summed score, but in a less direct manner. And providing only a 0 for the composite summed score, the *no response* carries less weight than a *correct* response, and much less than a

wrong one. Since it only occurs when neither *correct* nor *wrong* response is given, however, the *no response* nevertheless contributes to the sum score of each respondent.

This is partly because **premature responses** are recoded into a *no response*. When a response occurs less than 200 msec after the presentation of the task, it is classified as too quick for being an 'authentic' response to the stimulus. Consequently, it is automatically recoded into a *no response*. Interesting information may be lost in this recoding, however. While *no response* normally implies that processes in the respondent have been unacceptably *slow*, the premature responses are viewed as unacceptably *fast*. Using a common code for the rather opposite response types, then, serves to hide or suppress a potentially important difference. Being too quick and being too slow may well have opposite relations to executive functions, and the difference may not be trivial to our understanding of these functions. Could, e.g., an uncommonly short response delay be caused by an exceptionally agile working memory, while a delayed response is a signal of an active (and desirable) inhibition process?

The **response times** available in the Yellow/Red data may well also be related to more interesting problems than premature responses. Relating this 'delay' information to other research questions could prove to be a worthwhile effort. As we have just noted, inhibition would then be a likely candidate for attention. Another likely relation may be hypothesized between reaction time and item difficulty (Cf. paragraph 5.3). Do, e.g., response times covary with the difficulty of a task; and could they then be used for assessing the task's degree of difficulty? If so, would this assessment match the item difficulty scores of a Rasch analysis of the data?

It seems likely, however, that attention to the *wrong* response, the *no response* and the premature responses would be most useful in a clinical context, broadly speaking. The research aim then would be understanding the behavior of special groups and individuals, rather than describing main trends in larger parts of a population. If the focus is on deviance, differences, and unsolved problems, the 'less successful' responses may carry crucial information; and perhaps lead to insights that are relevant to remedial action, help and improvement. Finding ways of applying the Yellow/Red in this context is of course desirable; and further development of the games and their coding algorithms may contribute to a move in this direction.

For the present *Art of Learning* project, however, this may be less relevant. The focus of this project is executive processes with children *in general*, and no data of clinical relevance has been sought. The data material thus is not suited for explorations in a clinical direction. Also, strict Norwegian rules of research ethics will make it rather difficult to get acquire supplemental data of this kind. The idea of extending the use of 'non-correct' responses into a clinical context, therefore, will be left for others to pursue.

Hopefully, existing data sets elsewhere will include relevant information on deviant behavior, less successful education or personal shortcomings – in addition to Yellow/Red data. Exploring relations between clinically relevant data and 'non-correct' responses would then form a relevant challenge, and might even provide a basis for extending the usefulness of the Yellow/Red. This idea, however, will not be further pursued in the present report.

A couple of other ideas will be discussed, however, about the coding and information gathering from the Yellow/red games. The first implies *simpler coding rules*, only distinguishing between correct responses and all others. This may be compatible with giving equal weight to all tasks in a game, so that a general game score may be achieved by adding together the scores of all items.

In principle, his approach may also be extended into a Rasch model (Rasch, 1960; Wright & Mok, 2004), or perhaps an even more general IRT (Item Response Theory) scaling of the data (Nunnally & Bernstein, 1994). With this approach, separate 'difficulty scores' may be computed for each task in a game – in addition to deriving individual game scores for all informants. It may also allow for combining data from the four games into a more sophisticated general score for the entire Yellow/Red.

In practice, however, this would lead us far beyond the resource limits of the present report. Only the simple coding rules, therefore, will be further discussed here. Will they make a difference, or will they yield results that are similar the 'normal' coding procedures?

The second idea takes a simpler view of the *item differences*. It is based on the obvious differences of difficulty between the items, and on the naïve idea of assigning more points to the solving of harder tasks. With these assumptions, a more complicated coding algorithm will be called for. And again, the question will be whether it this makes any difference or not.

5.1 RECODING INTO CORRECT/NOT CORRECT RESPONSE

This recoding option simply makes a dichotomous separation between the *correct* responses and *all others*. One point (+1) is assigned to *correct* responses[3] and none (0) to the *not correct* (all others). By lumping 'all other responses' into the general category of *not correct*, a binary scale is achieved. This assumes that the most important information in the material is whether tasks/items were successfully managed or not. By implication, differences between the various *not correct* responses may be ignored. The new, recoded scores thus relate to a *simplified* binary scale.

This binary approach may have its advantages. If the above assumption of correct, the binary rescaling provides simpler and directly relevant data. Additionally, the number of cases in the data material remains high, since cases of *missing data* are not lost. They contribute to the *not correct* category of responses; and remain in all further computations. This is important to the Bindings game, as shown in paragraph 4.4.

Disadvantages are also likely, however. A central problem is the importance of the information that is lost through the binary recoding. The difference between *no response* and *wrong* ones, e.g., disappears. So does the distinction between 'normal' (level 1) and 'difficult' (level 2) tasks in the Bindings game. Assuming that some variance is related to these differences, the effect of this variance is uncertain but crucial. Does it mainly contribute to the understanding *correct* responses' influence in executive functions? Or is it rather part of the general 'noise' in the material, masking what the project intends to measure?

5.1.1 Correlating normal and binary scores

To get some idea of these questions, examining the relationship between the 'old' game variables and their binary counterparts may be useful. For an initial impression, correlations between the two measures were computed. This turned out to yield rather different results.

For *Cat/Dog* the correlation is 0.79, for *Trios* 0.96, for *Arrows* 0.93; and for *Bindings* 0.55. In the *Trios* and *Arrows* data, the correlations are very high, indicating a close correspondence. Quite likely, the old and the new binary variables both measure the same thing. The two versions are close to being interchangeable, so that preferring one over the other is difficult. For the *Cat/Dog* and the *Bindings* data, however, the correspondence is less impressive. This may be more interesting and calls for some explanation.

Scatterplots of the four old/binary relations may give some ideas about their differences. Figure 30 first displays the relationship between the old and the binary (recoded) data on the 33 tasks of the Cat/Dog game in the initial round of data gathering.

[3] Including all +1 codes, but also the +2 codes for getting difficult items right in the Bindings game.

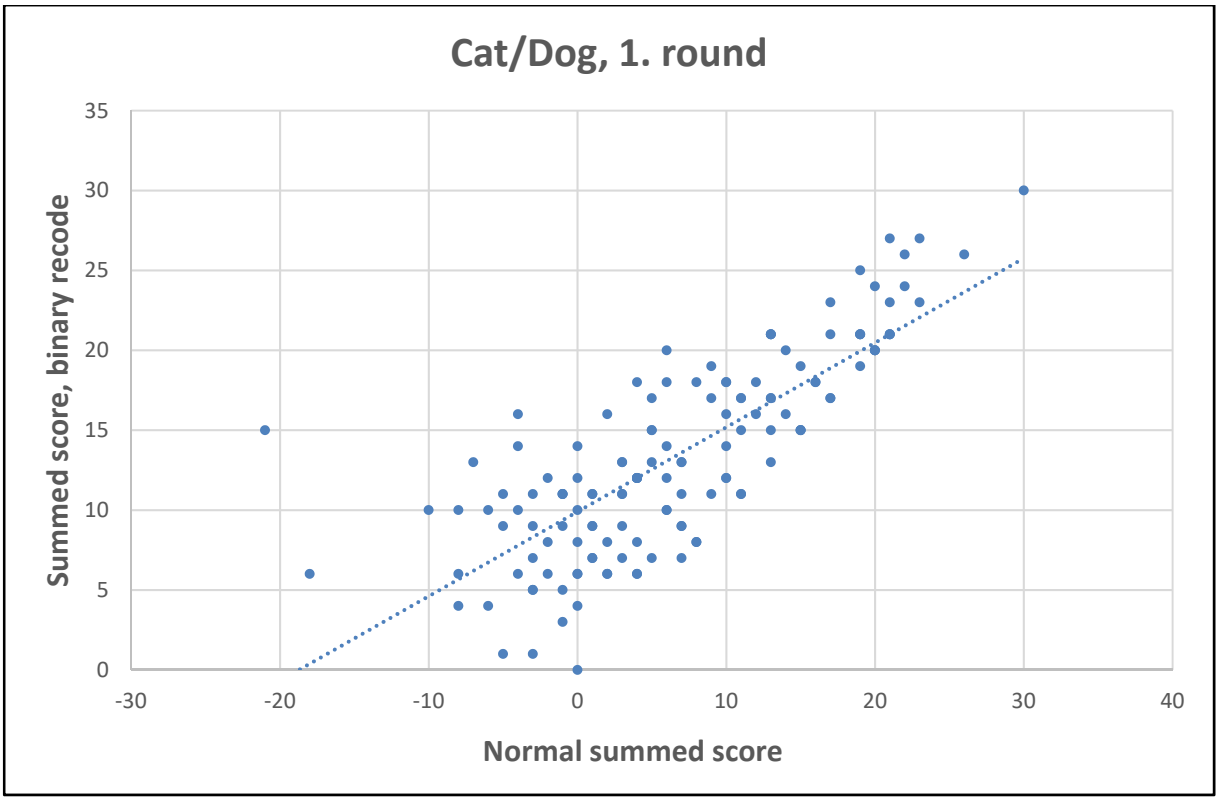

Figure 30: 'Normal' and binary sum scores in the Cat/Dog game of the first data round.

Firstly, the plot mirrors the clearly significant correlation ($r = 0{,}79$; $p < 0{,}001$) between the two different summed scores. All observations are found reasonably close to the trend line, confirming the close relationship of the variables. The two variables may well measure the same phenomenon.

At the same time, however, deviations from the line are also obvious. They indicate that the two variables are far from identical, but also appear to be influenced by some (non-trivial?) 'noise' or disturbance. There is unexplained variance in the data, suggesting that the trend line forms an insufficient model. There is more to be explained, and additional variables may be needed. Hence, the possibility of different pros and cons with the two coding procedures may not be ruled out.

Figures 31 and 32 on the next page show simpler patterns, however. In the Trios and Arrows games the relation between the 'old' and 'new' variables is very close. Generally, observations are situated closer to the trend line than in in Cat/Dog game. The 'normal' and the binary scales, therefore, basically measure the same thing. There also is limited 'noise' involved, so that different influences on the two variables appear unlikely. Consequently, there are no obvious arguments for preferring one coding paradigm over the other.

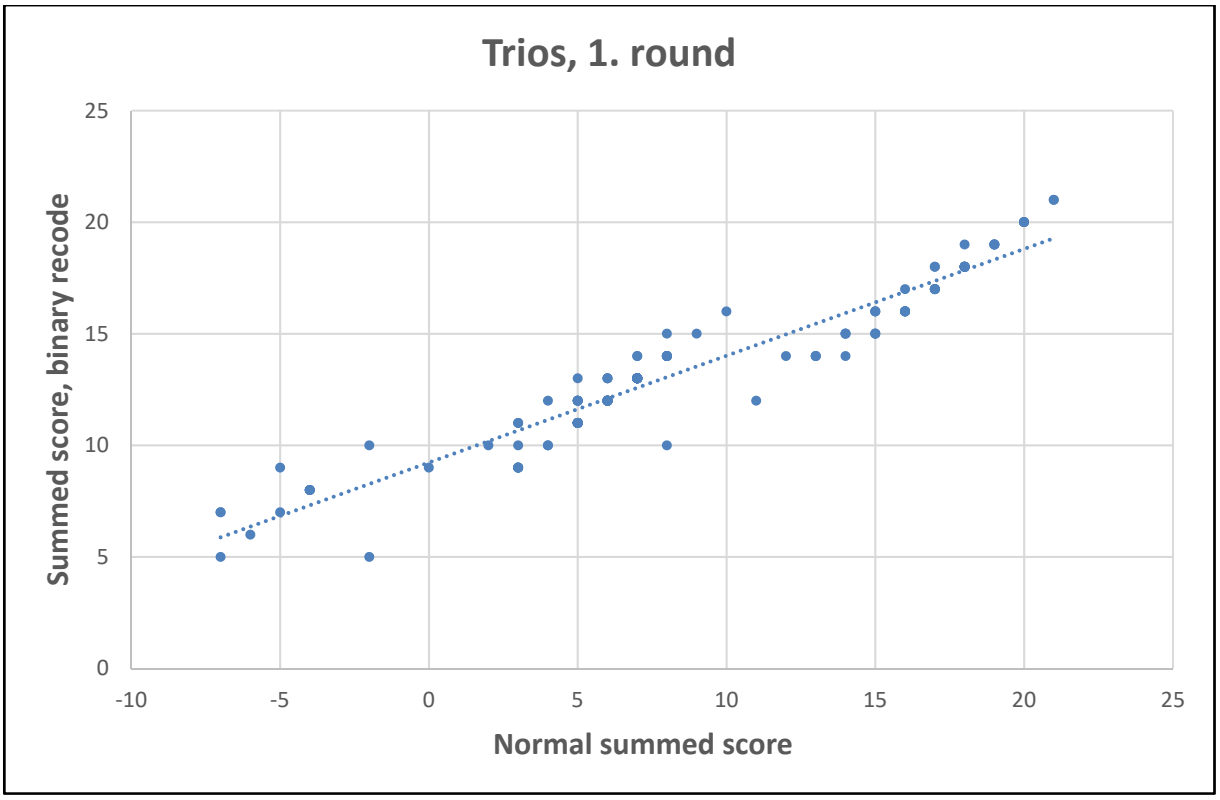

Figure 31: 'Normal' and binary summed scores in the Trios game of the first data round.

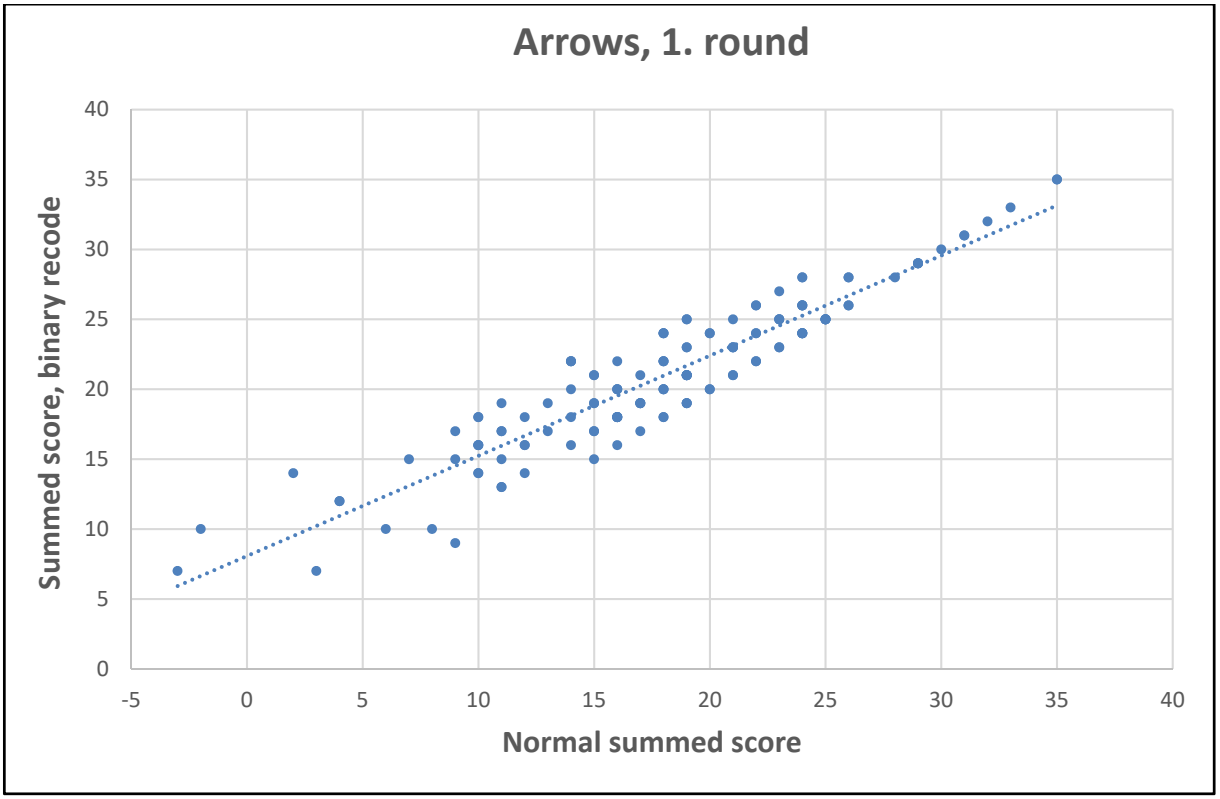

Figure 32: 'Normal' and binary sum scores in the Arrows game of the first data round.

When it comes to the Bindings game, however, it is again less simple. The correlation between the normal and the binary scales is considerably lower (0.55), even though it clearly is statistically significant ($p < 0.001$).

And figure 33 shows that the observations are not generally close to the trend line. Many observations clearly deviate from it, especially those with high and low scores on the old-fashioned summed score. Using some imagination, the plot may even suggest a curvilinear relationship.

In the lower parts of the scale, the correspondence between rising values on both scales is no surprise. But at high normal scale values, increasing values apparently are accompanied by *decreasing* values on the binary scale. This is hardly an expected result.

It may thus deserve some attention. Is this simply due to random noise, or does it signal some more substantial influence that is not understood? If so, an obvious next step might be to see if these relationships also appear in the second and third rounds of data procurement. The present report, however, will not explore these questions any further.

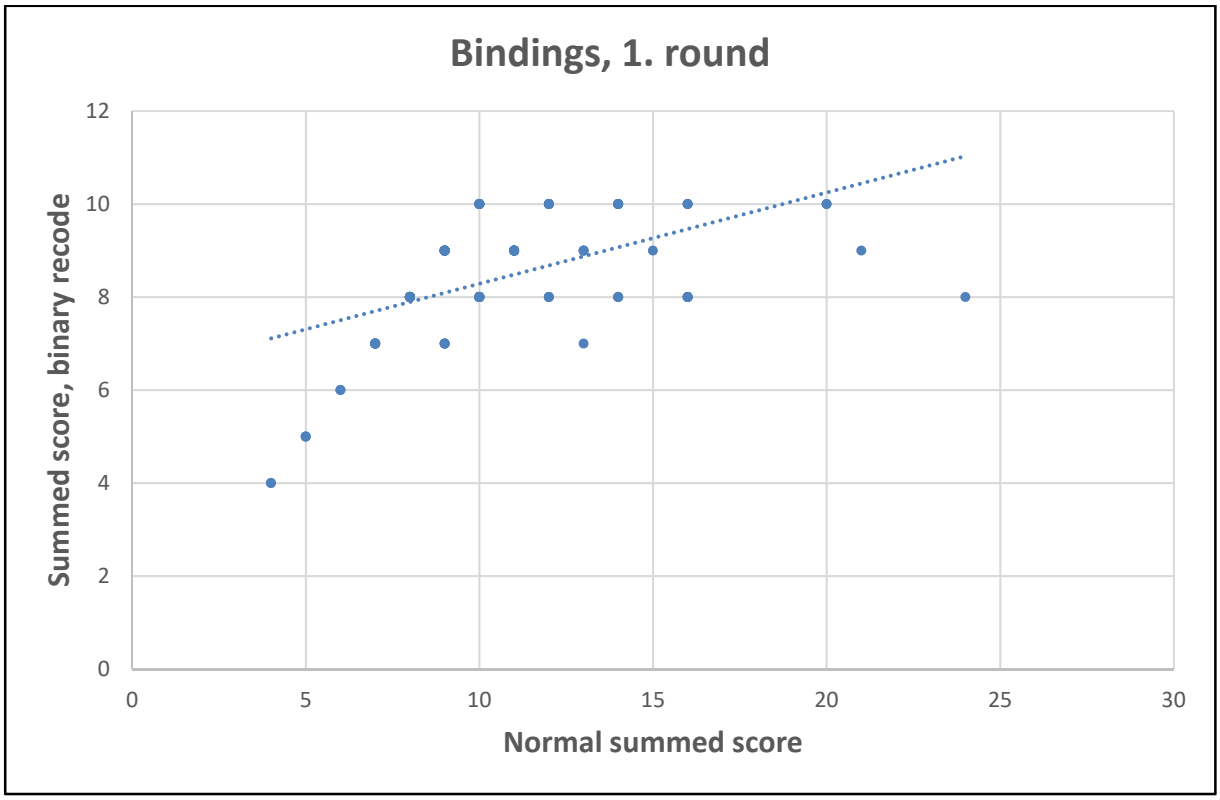

Figure 33: 'Normal' and binary sum scores in the Bindings game of the first data round.

So far, however, we have seen that recoding into binary scores does not matter much to the Trios and the Arrows data. With these two scales, the 'normal' and the binary coding schemes produce quite similar results. Data from the Dog/Cat and the Bindings games, however, appear to change more with the binary recoding; and the two coding schemes do not produce the same result.

A quick look at figures 14 – 17 may suggest that the *wrong* responses are less frequent in the Trios and Arrows games. The pattern of *not correct* responses thus is a bit less complex, yielding less variance than what is seen with the Cat/Dog and Bindings games. This may partly explain why the binary recoding involves less change of the Trios and the Arrows data than it does with Cat/Dog and Bindings.

5.1.2 Measurement properties of simplified scores

The new summed scales appear to have acceptable measurement properties. Their distributions are shown in Table 7. Please note that all informants are now included in the data (Cf. paragraph 2.1 and Table 2).

Table 7: Basic statistics of the four simplified scales, first data round (N=139)

	Cat/Dog	Trios	Arrows	Bindings
Mean	13.4	13.4	21.1	9.2
Median	12.0	13.0	21.0	9.0
Mode	11.0	13.0	18.0*	8.0
Standard deviation	6.1	3.4	5.4	2.0
Minimum	0.0	5.0	7,0	4.0
Maximum	30.0	21.0	35.0	16.0
Number of items	33	21	36	27

* Two modes exist (18 and 24). The smallest value is shown.

In figures 34 – 37, the distributions of the "new" game scales are displayed. They all come close to forming normal, bell-shaped, and symmetrical curves. Data with these properties will normally be preferred for statistical procedures that assume normal data distributions. With this consideration in mind, some quick comparisons to figures 2, 4, 6, and 8 may be in order.

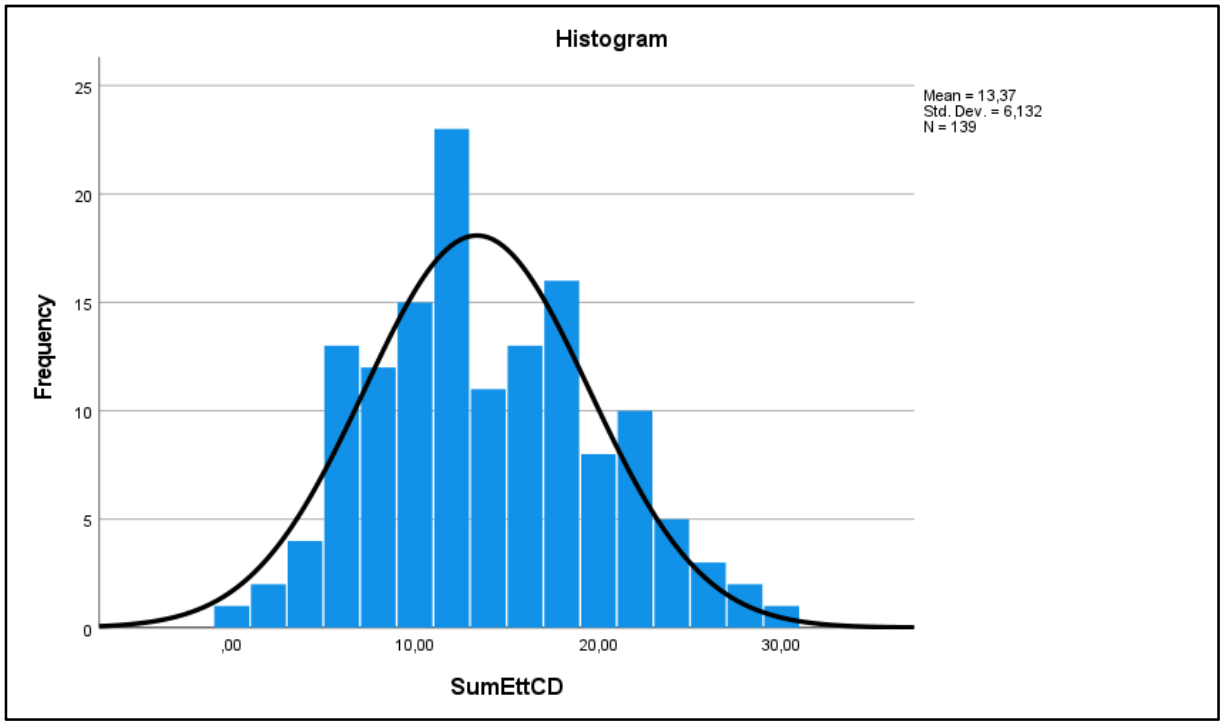

Figure 34: Score distribution of simplified Cat/Dog scale, first data round

At first sight, distributions of the 'old' scales here appear slightly less 'normal', often containing more 'outliers' than the 'new' ones. This informal impression may thus suggest that the new 'binary' scales have some advantages over the traditional scores of the Yellow/Red, and thus are worthy of some attention.

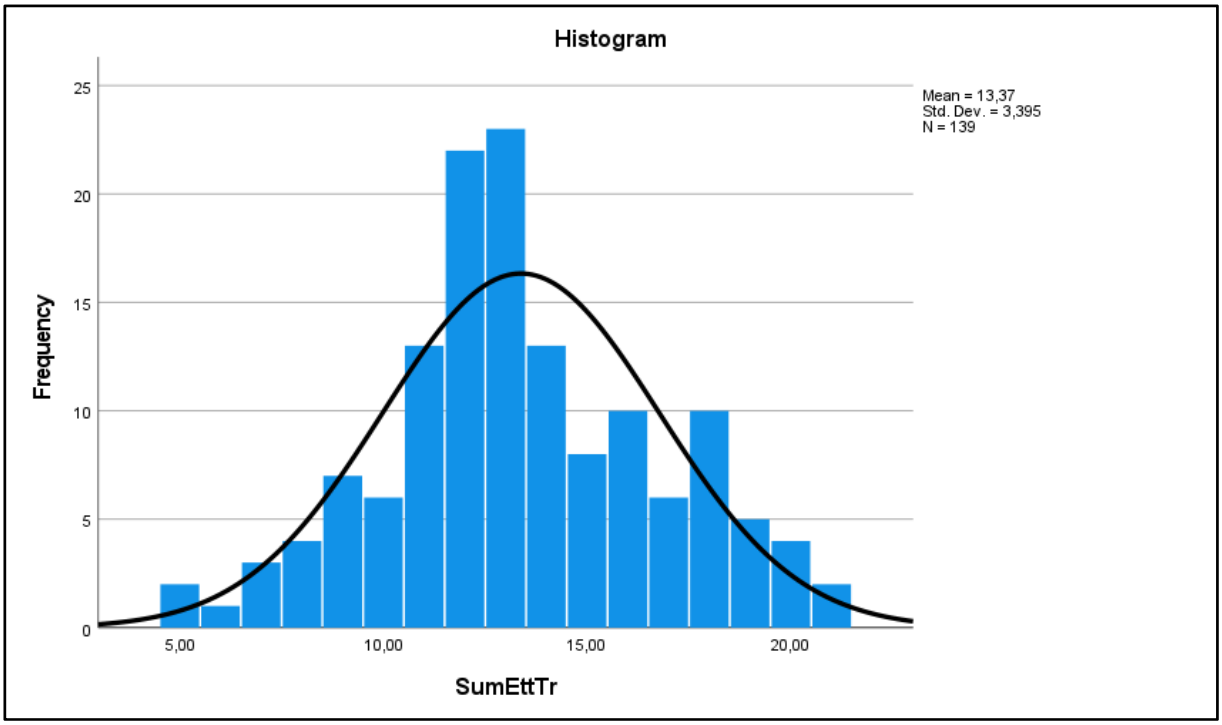

Figure 35: Score distribution of simplified Trios scale, first data round

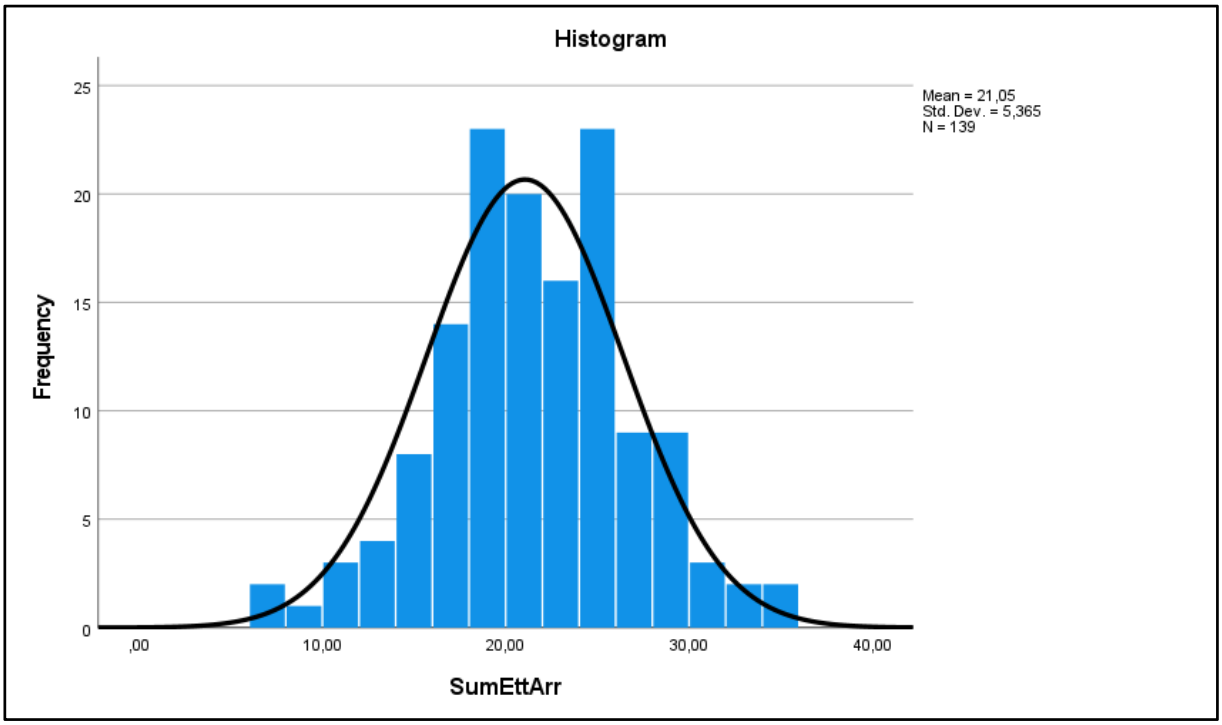

Figure 36: Score distribution of simplified Arrows scale, first data round

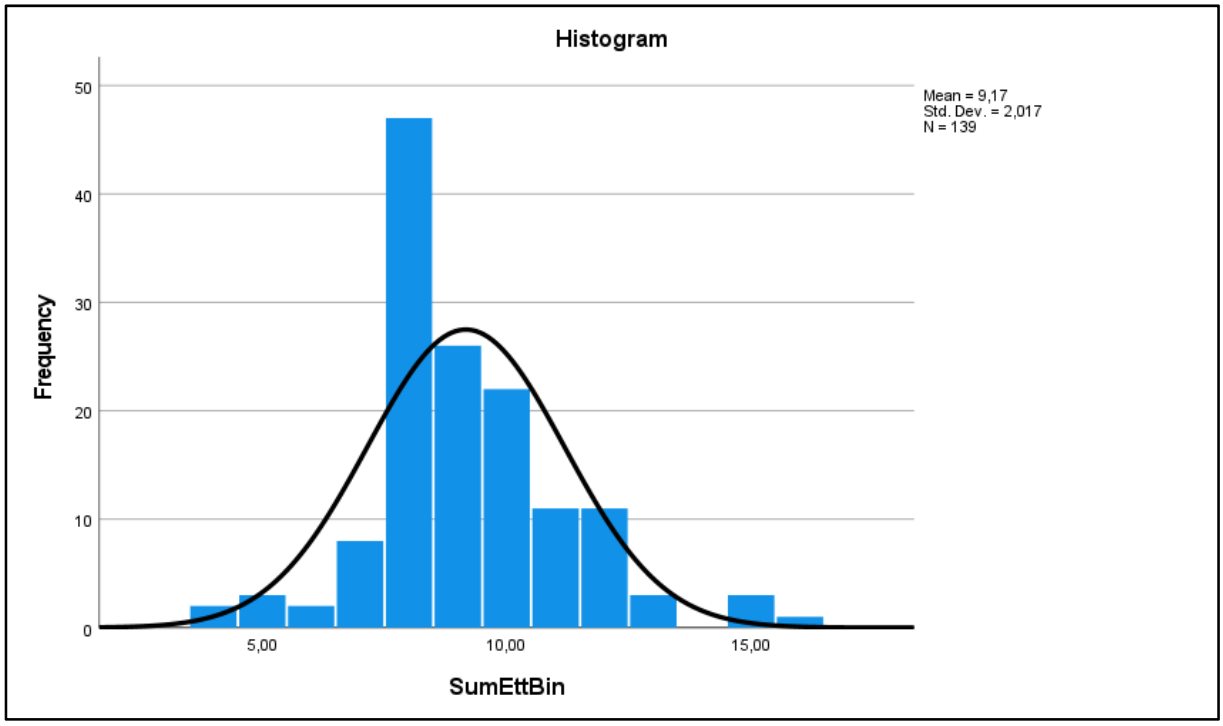

Figure 37: Score distribution of simplified Bindings scale, first data round

Please note, however, that the simplified scales provide a much larger sample (N = 139) than do the traditional measurements of the Yellow/Red. Including *no response* as well as *missing data* responses in the general 'not correct' category, the new coding rules manage to retain all respondents in the sample. The two sets of scales thus have different samples, and comparisons should only be done with considerable caution.

5.1.2.1 Reliability of simplified scores

In table 8, Cronbach's *alphas* of the simplified scales are shown for all three waves of data gathering. Comparing this to table 2 is not quite straightforward, but rather interesting differences appear.

Table 8: Cronbach's alpha (and N) for binary scales of four games at three time points

	Pre-intervention	**Just after int.**	**Six months later**
Cat/Dog (33 items)	0.83 (N = 139)	0.85 (N =139)	0.83 (N = 139)
Trios (21 items)	0.75 (N = 139)	0.81 (N = 139)	0.78 (N = 139)
Arrows (36 items)	0.82 (N = 139)	0.85 (N = 139)	0.82 (N = 139)
Bindings (27 items)	0.64 (N = 139)	0.76 (N = 139)	0.70 (N = 139)

For the Cat/Dog game, the alphas are high throughout; the rescaling does not change the *alphas* very much. For Trios, however, the alphas of binary scales are clearly higher than with the traditional scales. Also, the N of the computations are notably increased. There also is some improvement with the Arrows data, but more moderate. And here, N was high already with the 'old' version of the scale.

For the Bindings game, it should first be noted that computing alpha for this scale still has its problems. At the initial time point, *all* respondents had a *correct* response to item 2 and *none* to items 23 through 27. These 6 items thus have no variance on the binary scale and were removed from the computations of *alpha*. In the second data and third data waves, however, the number of zero-variance items had risen to 10, meaning that only 17 items could be meaningfully included in the *alpha* computations.

Even after this removal of items, however, computing *alphas* for the simplified scales was possible. Moreover, he alphas for Bindings even have interesting values across the time span, suggesting rather consistent measurements.

Generally, then, *alphas* of the binary scales are *not* inferior to the original scales. The are about the same for the Cat/Dog scales, improved for Trios and Arrows, and they are (at least) computable for Bindings. The reliability thus is more convincingly shown with the simplified scales, supporting our continued interest in the new binary coding. It should be kept in mind, however, that the improvements are related to increased sample size.

The reliability of the total summed Yellow/Red scores may also be of interest. It is computed by adding the summed scores of all four binary Yellow/Red scales. The *alphas* at the three time points are shown in table 9 on the following page.

The reliability of the binary or simplified version of the all-over Yellow/Red score is only acceptable for research purposes. The slight improvement over the values shown in table 3 is not convincing. It may be noted, however, that neither the alphas of 'traditional' scales nor the alphas of the binary versions change much over time.

Table 9: Cronbach's alpha for the sum of all binary scales at three time points

	Pre-intervention	**Just after int.**	**Six months later**
Yellow/Red summed	**.65**	**.63**	**.65**
Excluding Cat/Dog	.62	.55	.50
Excluding Trios	.53	.57	.60
Excluding Arrows	.44	.50	.46
Excluding Bindings	.66	.62	.66

As shown in the article of Andersen et al. (2019, table 4), however, also BRIEF indices change over time. This should be kept in mind, since it makes the use of BRIEF data for validating the Yellow/Red scales less straightforward.

5.1.2.2 Validity of simplified scores

As shown in chapter 2, support for the validity of the four main Yellow/Red scales is also not easily found. In the present data set, only the BRIEF data contain independent information on executive

processes that may be used for validation. Consequently, the BRIEF data should also be tried for validating the simplified Yellow/Red scores.

Table 10: Correlations between 11 BRIEF indices and four binary scales, first data round

BRIEF index		Cat/Dog	Trios	Arrows	Bindings
Inhibit N=135	Pearson r	-0,03	-0,16	-0,06	-0,03
	Sig. (2-tailed)	0,70	0,06	0,49	0,72
Shift N=135	Pearson r	-0,10	-0,13	-0,09	-0,12
	Sig. (2-tailed)	0,24	0,14	0,32	0,17
Emotional control N=136	Pearson r	0,02	-0,10	-0,09	-0,01
	Sig. (2-tailed)	0,79	0,26	0,31	0,92
Initiate N=136	Pearson r	-0,08	-0,19	-0,20	-0,14
	Sig. (2-tailed)	0,38	0,03	0,02	0,10
Working Memory N=134	Pearson r	-0,10	-0,31	-0,21	-0,14
	Sig. (2-tailed)	0,27	0,00	0,01	0,10
Plan/organize N=136	Pearson r	-0,03	-0,17	-0,04	-0,12
	Sig. (2-tailed)	0,71	0,05	0,61	0,16
Organize Materials N=126	Pearson r	-0,10	-0,18	-0,13	-0,04
	Sig. (2-tailed)	0,26	0,04	0,14	0,64
Monitor N=138	Pearson r	0,00	-0,17	-0,07	-0,06
	Sig. (2-tailed)	0,96	0,04	0,41	0,48
Behavior regulation N=130	Pearson r	-0,05	-0,15	-0,08	-0,06
	Sig. (2-tailed)	0,61	0,08	0,34	0,49
Metacognition N=120	Pearson r	-0,07	-0,25	-0,17	-0,13
	Sig. (2-tailed)	0,43	0,01	0,07	0,16
Global executive N=112	Pearson r	-0,07	-0,22	-0,12	-0,11
	Sig. (2-tailed)	0,43	0,02	0,21	0,23

In table 10, the correlations from the initial data wave are shown. Generally, the match between the BRIEF and the Yellow/Red data are less than impressive. The highest number of significant inter-correlations (7) is seen with the Trios scale; suggesting that it does in fact measure some parts of what is covered by the BRIEF. As shown in table 1, however, the Trios game was really expected to tap *cognitive flexibility*. This also appears to be the idea behind the BRIEF index of *Shift*. Unfortunately, the two are not significantly correlated in table 10.

A comparison with table 4, however, shows some differences. Here, the highest number (5) of intercorrelations was with the Arrows scale, not with the Trios. But some similarities also appear. Indices *Initiate* and *Working Memory* also correlate with the Arrows scale in table 10. The significant correlation between Trios and *Working Memory* is also found in both tables. Moreover, for both scales several correlations in the two tables come close to being significant; indicating that the difference between the tables is not an absolute one.

The patterns have changed in the second data round, as shown in Table 11. The Arrows scale now has become the important predictor, being significantly correlated to six of the eleven BRIEF indices. The same change occurred with the 'old' scales, in tables 5 and 6.

Table 11: Correlations between 11 BRIEF indices and four binary scales, second round

BRIEF index		Cat/Dog	Trios	Arrows	Bindings
Inhibit N=120	Pearson r.	0,00	-0,07	-0,17	-0,02
	Sig. (2-tailed)	0,99	0,44	0,07	0,79
Shift N=121	Pearson r.	-0,01	-0,06	-0,12	-0,06
	Sig. (2-tailed)	0,91	0,53	0,21	0,48
Emotional control N=123	Pearson r.	0,06	0,03	-0,05	-0,02
	Sig. (2-tailed)	0,52	0,75	0,59	0,85
Initiate N=124	Pearson r.	-0,12	-0,16	-0,26	-0,09
	Sig. (2-tailed)	0,17	0,08	0,00	0,31
Working Memory N=125	Pearson r.	-0,14	-0,18	-0,32	-0,09
	Sig. (2-tailed)	0,11	0,04	0,00	0,32
Plan/organize N=121	Pearson r.	-0,13	-0,05	-0,18	-0,03
	Sig. (2-tailed)	0,15	0,61	0,04	0,74
Organize materials N=120	Pearson r.	0,04	0,06	-0,11	-0,03
	Sig. (2-tailed)	0,68	0,48	0,23	0,76
Monitor N=122	Pearson r.	0,00	-0,07	-0,19	-0,04
	Sig. (2-tailed)	1,00	0,47	0,04	0,63
Behavior regulation N=117	Pearson r.	0,02	-0,02	-0,14	-0,04
	Sig. (2-tailed)	0,81	0,82	0,15	0,68
Metacognition N=113	Pearson r.	-0,09	-0,09	-0,25	-0,06
	Sig. (2-tailed)	0,35	0,33	0,01	0,50
Global Executive N=109	Pearson r.	-0,04	-0,06	-0,21	-0,06
	Sig. (2-tailed)	0,70	0,53	0,03	0,55

Even more changes are evident in the third data round, however. Finally, the Cat/Dog scale shows some predictive power, significantly correlating to six of the eleven BRIEF indices. The Arrows scale also appears as an interesting predictor, with significant correlations to all BRIEF indices but one (Organize materials). At long last, the Arrows now also is significantly related to the *Inhibit* index, as theory would predict.

This all-over pattern is strikingly like that of table 6. Generally, then, validity of the binary version and the original versions of the Yellow/Red scales appear to be very similar – as assessed by their relations to the eleven BRIEF indices.

Moreover, the criterion validity (as assessed through BRIEF indices) of *both types of scales* changes considerably over three points of time. Initially, mainly the Trios scale is related to the BRIEF indices. At the next step, however, the Arrows scale appears to provide the strongest relationship to the BRIEF. Finally, however, both Arrows and Cat/Dog have much in common with most BRIEF indices.

All in all, however, the eleven BRIEF indices do not appear to be useful validators of the four Yellow/Red scales – neither in their original nor in their simplified versions. No other variables for alternative validation are found in the present data set. The validity of the Yellow/Red scales, therefore, is not proven in this project.

Table 12: Correlations between 11 BRIEF indices and four binary scales, third round

BRIEF index		Cat/Dog	Trios	Arrows	Bindings
Inhibit N=97	Pearson r.	-0,15	-0,02	-0,29	-0,12
	Sig. (2-tailed)	0,13	0,86	0,00	0,23
Shift N=99	Pearson r.	-0,24	-0,09	-0,30	-0,11
	Sig. (2-tailed)	0,02	0,38	0,00	0,28
Emotional control N=96	Pearson r.	-0,10	-0,05	-0,23	-0,18
	Sig. (2-tailed)	0,32	0,64	0,03	0,09
Initiate N=97	Pearson r.	-0,25	-0,01	-0,27	-0,12
	Sig. (2-tailed)	0,01	0,90	0,01	0,26
Working memory N=99	Pearson r.	-0,31	-0,12	-0,32	-0,17
	Sig. (2-tailed)	0,00	0,25	0,00	0,10
Plan/Organize N=97	Pearson r.	-0,29	0,07	-0,21	-0,12
	Sig. (2-tailed)	0,00	0,47	0,04	0,24
Organize materials N=98	Pearson r.	-0,11	0,10	-0,09	-0,11
	Sig. (2-tailed)	0,29	0,32	0,36	0,28
Monitor N=99	Pearson r.	-0,24	-0,01	-0,34	-0,18
	Sig. (2-tailed)	0,02	0,94	0,00	0,08
Behavior regulation N=93	Pearson r.	-0,17	-0,04	-0,31	-0,16
	Sig. (2-tailed)	0,11	0,71	0,00	0,12
Metacognition N=92	Pearson r.	-0,24	0,04	-0,29	-0,14
	Sig. (2-tailed)	0,02	0,73	0,01	0,18
Glocal executive N=88	Pearson r.	-0,21	0,01	-0,32	-0,16
	Sig. (2-tailed)	0,05	0,90	0,00	0,14

5.1.2.3 The usefulness of simplified scores

The central question is: Are the simplified scores any good? Are they in any way better than the original scales, and should they be preferred for any reason?

As shown in paragraph 5.1.1, the simplified Trios and Arrows scales are hardly distinguishable from the original scales. Very high correlations indicate that simplified and original scales mainly measure the same things, making them virtually interchangeable. Preferring one over the other is hardly important.

For the Cat/Dog and the Bindings games, however, the picture is less clear. While the original and the simplified scales are significantly correlated, the correlations do not explain all the variance. Additional variance does exist, and deviations from the regression line are obvious for both games. This clearly leaves room for the simplified scales to yield results that are not identical to those produced by the original scales.

Neither reliability nor validity seems to be improved in the simplified scale versions, as shown in the two previous paragraphs. An analysis of variance was also performed to estimate the effects on the simplified scores from the repeated measures and the intervention/control group difference. The results were the same as those reported in chapter 4.

All in all, therefore, the simplified binary scales do not appear to be any better than the original scales. However indirectly, this gives some support for the more complex coding algorithms of the four Yellow/Red scales.

5.2 RECODING BY ITEM DIFFERENCES

Chapter 4 clearly shows that differences between items are far from trivial. Although the set of respondents is relatively constant, items in all four games elicit quite a range of different responses. But by treating all items the same way, the traditional summed scores may serve to 'hide' potentially interesting variance. To further improve our understanding of the Yellow/Red test, therefore, we should look beyond the traditional summed scores. Attention should be directed to the obvious differences between the many items involved. It may be useful, however, to distinguish between two kinds of item differences.

The first type of difference is simply variations in item difficulty. Certain items yield correct responses from most respondents and may be viewed as 'easy'. Other items rarely elicit correct responses, and thus are 'difficult'.

The second kind of difference is the task or the focus of the items. The four Yellow/Red games are intended to assess different aspects of executive functions, like inhibition, cognitive flexibility and working memory, as sketched in table 1. These aspects, however, may also indicate distinctions between the items within a game. In the Cat/Dog game, e.g., some items may primarily activate working memory, while a successful response to other items also assumes some inhibition. This type of difference does not necessarily coincide with variations in difficulty.

5.2.1 Item difficulty variation

A common approach to dealing with variations of item *difficulty* is Rasch scaling (Bond & Fox, 2007; Wu & Adams, 2007). This may clearly be an option also for analyzing the present material but is rather complicated in practical use.

There may be easier procedures, however, for including item difficulty in the data analysis. Simply scoring more points for solving more difficult items, and less for solving the easier ones may be a useful start. This general approach is already used with the Bindings game, where items were constructed to fit three different levels of difficulty and are scored accordingly.

The general idea may be extended to all four games, however. First, a 'difficulty' score for each item is needed. For this, the percentage of respondents *not* yielding the correct response may be a practicable index. The potential range of this runs from 0 (Easy: *all* respondents give *correct* responses) to 100 (Difficult: *no* respondents are *correct*). For each item, this index is assigned as the score for all respondents giving the 'correct' response, while all others receive a 0. In principle, the resulting data pattern may then include more information than the binary scores with equal weight to all correct responses.

5.2.1.1 The Cat/Dog game

This algorithm was first applied to the Cat/Dog scores of the first data round. The range of the 33 item indices from this game runs from 27 to 79, with a mean of 59.

The mean of the resulting 33 scores for each person then becomes the respondent's score for the game. For the Cat/Dog game, these scores range between 0 and 53, with a mean of 22. A graph of

the resulting set of scores is displayed in figure 38. It does not look very different from a bell-shaped 'normal' curve.

However, these mean difficulty scores turn out to be perfectly correlated (r = 0.99) with the simplified binary scores. In practical terms, these scores are thus interchangeable, and one may hardly be preferred before another.

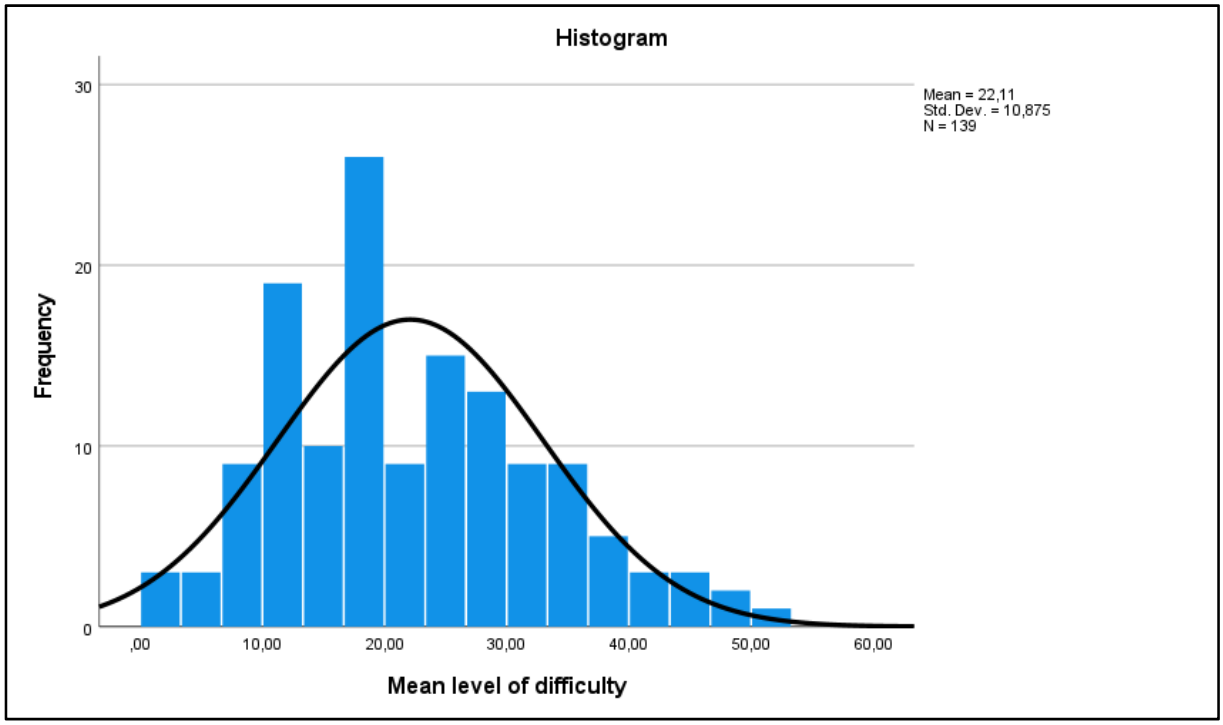

Figure 38: Score distribution of difficulty-weighted Cat/Dog scale, first data round

5.2.1.2 The Trios Game

Applying the same procedures to the Trios data, however, produces a rather different result. Here, the range of the item indices fall between 2 and 36, with a mean of 16. This game thus clearly is more difficult than the Cat/Dog game ($t = 6.400$; $df = 138$; $p < 0.001$).

The graph in figure 39 on the following page also is rather different. It seems to suggest a bimodal distribution of the average difficulty scores, not a normal distribution. Looking back at figure 35, it offers no hint of this apparent partition of the respondents. Evidently, therefore, these average difficulty scores do carry some information in addition to what is covered by the bimodal summed scores.

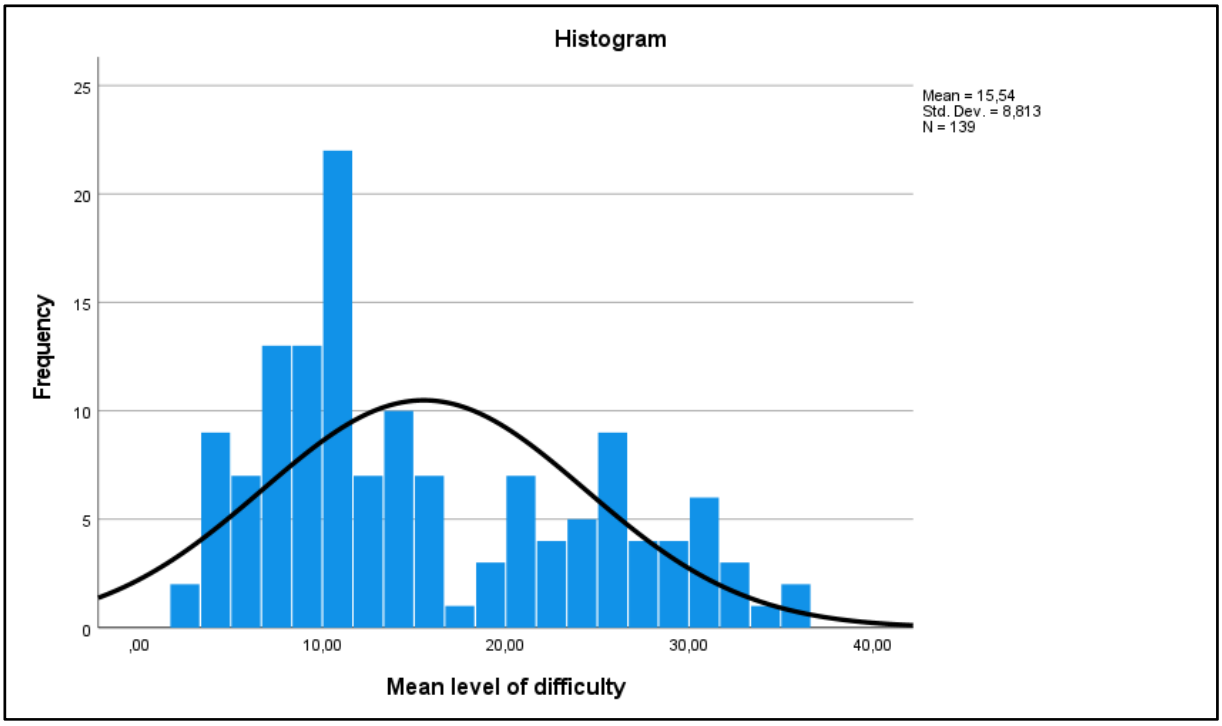

Figure 39: Score distribution of difficulty-weighted Trios scale, first data round

Looking back at figure 21, however, may provide a relevant clue. Here, item TR11 stands out as the most difficult item by far in this game, producing a disproportional number of errors. In the first data round, only 29 of the 139 respondents respond correctly to it.

This problem also extends to items TR12 through TR15, where large (and unexpected) numbers of missing data are produced. Consequently, TR11 appears as a major predictor of the response to item TR12, as shown in table 13. A correct response to TR11 is normally (86%) followed by a correct response to TR12. And vice versa: 'Other' (non-correct) responses to TR11 most often (77%) means some 'other' response also to TR12. This bias is statistically significant (Chi-square = 40.153; df = 1; $p < 0.001$). Moreover, it is also reproduced in the relationships between TR11 and items TR13 through TR15.

Table 13: Correct and 'other' responses to items TR11 and TR12.

	Correct, TR12	**Others, TR12**	**Total**
Correct, TR11	25 (86%)	4 (14%)	29 (100%)
Others, TR11	25 (23%)	85 (77%)	110 (100%)
Total	50 (36%)	89 (64%)	139(199%)

It is rather interesting, therefore, to look at the effect of the TR11 on the difficulty-weighted Triads score. For respondents with success on the TR11, the mean of the difficulty-weighted score is 26.3; while it is only 12.7 for respondents with 'other' responses. This difference is clearly significant (t = 9.545; df = 137; p < 0.001). It may thus explain the bimodality shown in figure 37, at least partly. The correlation between the summed difficulty scores and their bimodal counterparts is very high (r = 0.93) even here, however.

The summed difficulty scores are also clearly related to the subjects' age. After including the single 9-year-old in the 8-year group, there are three age groups in the material. The three groups have different difficulty scores, and the mean difference is clearly significant (F = 4.817; df = 2; p = 0.01). Comparable age differences also exist with the original Triad measure (F = 5.574; df = 2; p = 0.005) as well as the corresponding binary scores (F = 5.187; df = 2; p = 0.007).[4]

However clear, this age difference is no explanation of the bimodal distribution of figure 39. As shown in figure 40 on the following page, age groups and TR11 responses are independent influences in the difficulty-weighted Triads score. This is confirmed by a univariate ANOVA. It shows not only that the TR11 response is a significant effect (F = 63,966; p < 0.001). Age is also significant (F= 3,450; p = 0,035), while the interaction effect is not (F = 0.270; p = 0.764). It may also be noted that the regression model implied in this ANOVA is rather strong (R^2 = 0.448).

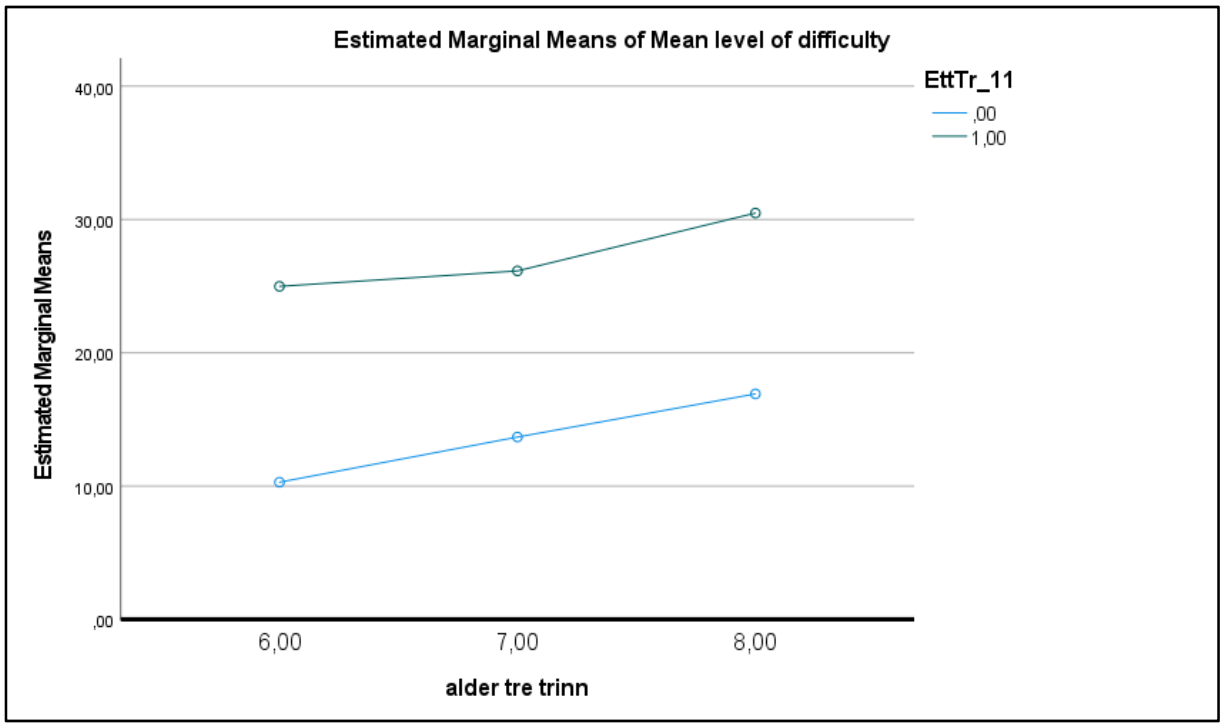

Figure 40: Mean difficulty scores of Trios scale in age groups vs. correct and 'other' TR11 responses

[4] Neither should be confused with the changes over time shown in chapter 3, where the ANOVA includes the changes over time in all age groups.

5.2.1.3 The Arrows game

The Arrows data may also suggest a bimodal distribution, as shown in figure 41. The mean of the individual difficulty scores here is 16, and they range from 4 to 39. This is significantly more difficult than the Cat/Dog data ($t = 7.547$; $df = 138$; $p < 0.001$). The general levels of difficulty for the Arrows and the Trios data, however, are not different.

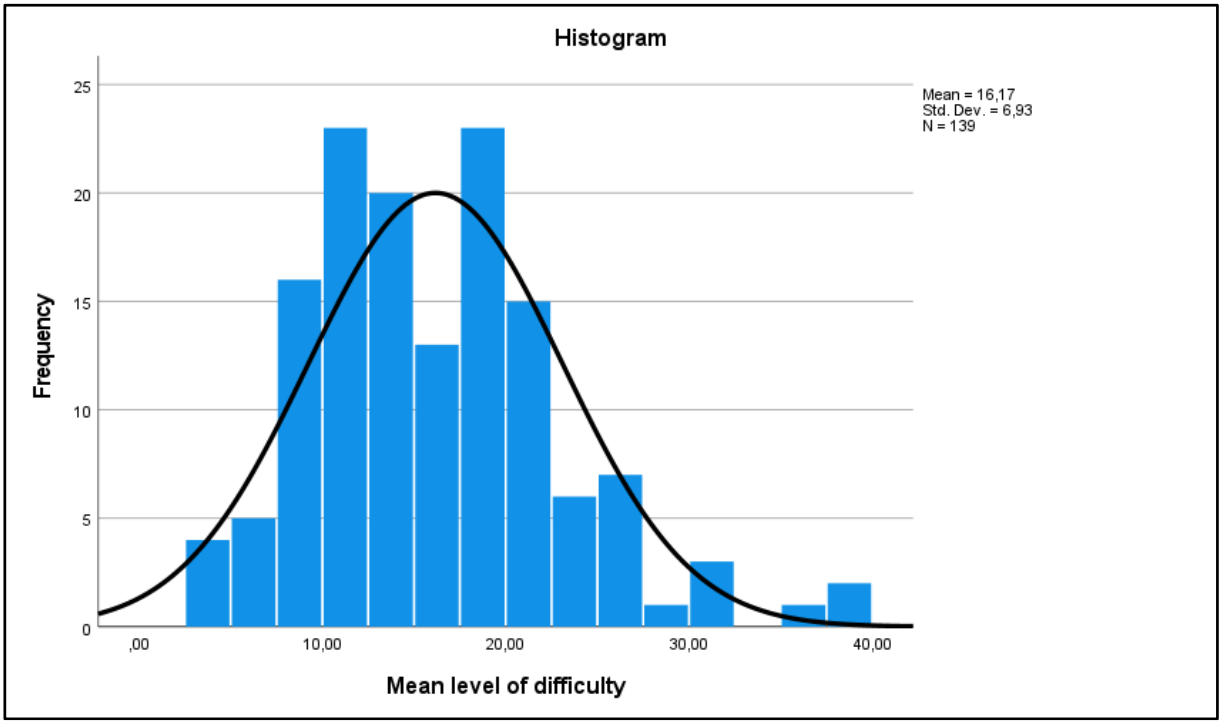

Figure 41: Score distribution of difficulty-weighted Arrows scale, first data round

5.2.1.4 The Bindings game

The Bindings game, however, yields quite another picture; as shown by figure 42. It appears as unimodal, and its distribution of scores is heavily skewed towards the minimum. Also, it is by far the most difficult of the four games.

Its individual scores range from 0 (or more precisely 0.15) to 29 and have a mean of 5.8. This is considerably lower than both Trios ($t = 12.383$; $df = 138$; $p < 0.001$) and Arrows ($t = 15.546$; $df = 138$; $p < 0.001$). Clearly, this game is too difficult for the young children in our sample; and may thus not be suited for making distinctions between them.

Even here, the difficulty score's correlation with its binary counterpart is very high at 0.94. The better part of the variance, evidently, is common to the two variables. Preferring one over the other in further work, therefore, is not an easy decision.

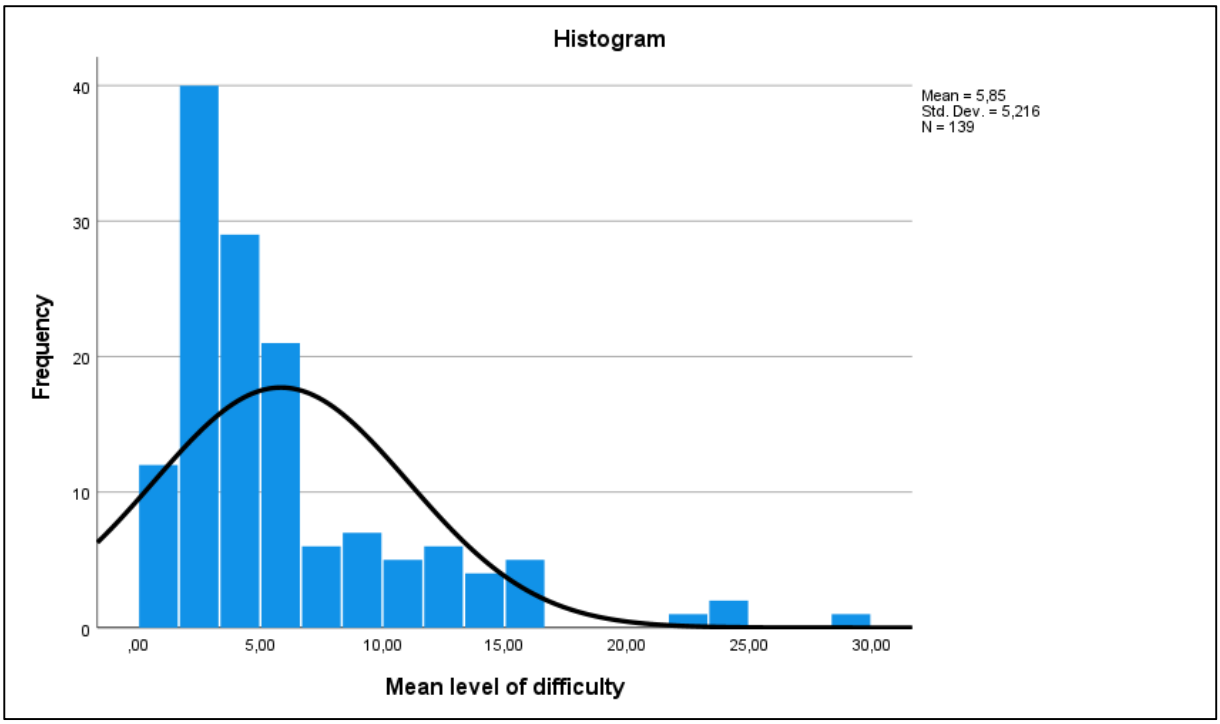

Figure 42: Score distribution of difficulty-weighted Bindings scale, first data round

The obvious similarity between the difficulty-weighted and the 'simplified' binary scores also comes out when reliability and validity are considered. The Cronbach *alpha* values of the difficulty-weighted Cat/Dog and Triads scales are at .82 and .77, respectively. Corresponding figures for Arrows and Bindings are .80 and .67. All numbers are very similar to the pre-intervention alphas of the binary versions for the four game scales, as shown in table 8. Comparing BRIEF/difficulty-weighted correlations to those shown in table 9 also reveal no interesting differences.

All in all, therefore, our simple recoding of the 'simplified' scores into difficulty-weighted scores is no obvious success. The new scales do not offer much new information. Generally, the correlations between the two sets of scores are extremely high, and little is gained by substituting one for the other. Consequently, further work on the difficulty-weighted scores is not seen as appropriate. The unexpected bimodal distribution of the difficulty-weighted scores, however, may suggest that something *not understood* may be hiding in the data.

5.2.1.5 Comparing the four games

What the difficulty-weighted scores show clearly, however, is a definite difference between the four games. As shown in figure 43 on the following page, the Cat/Dog game has the highest average score, while the Bindings game is ay the opposite end. The means of the Trios and Arrows games fall between these two extremes. A repeated-measures ANOVA show that the mean differences between the four games is statistically significant (MS = 6298.228; F = 129.205; $p < 0.001$).

This does *not* mean, however, that The Cat/Dog game is the most difficult of the four. While containing a comparatively large number of easy items, it also has a high number of 'correct' responses. Since relatively few 0-scores therefore are involved, the average score is high. The high 'hit' rate more than compensates for the low 'pay-off' of the 'correct' responses. With the Bindings game, it is the other way around. Very few 'right' responses are given, and the high score of the many

difficult items can not compensate for the many 0-scores. Somewhat paradoxically, the easiest game thus produces the highest mean score on the difficulty-weighted scale.

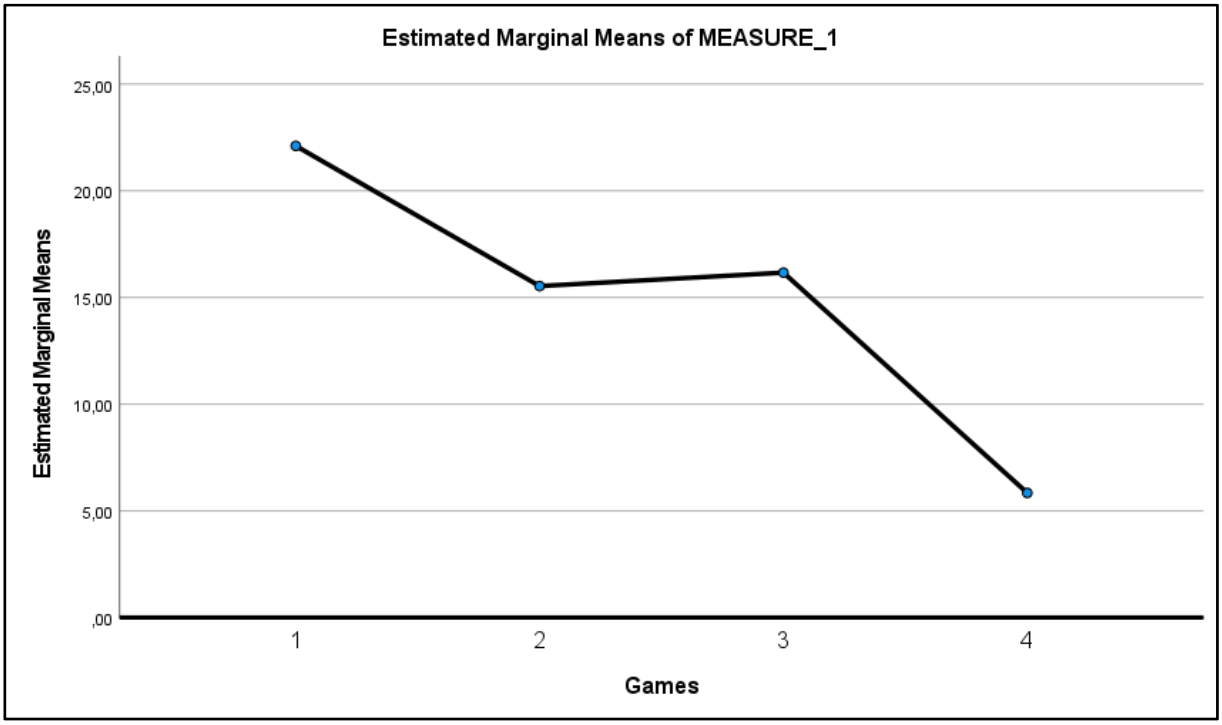

Figure 43: Mean of difficulty-weighted respondent scores in four games, first data round (N=139).

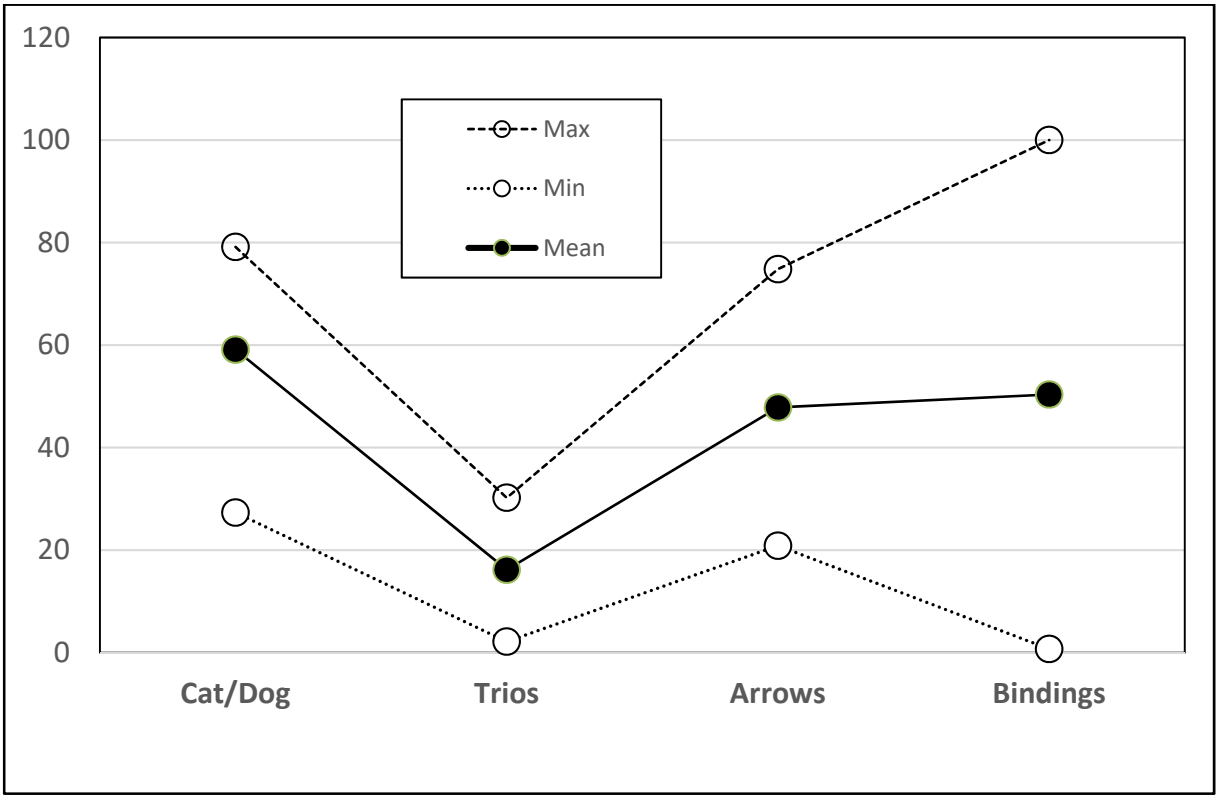

Figure 44: Item difficulty of the four games. Mean, minimum and maximum scores.

As shown in figure 44, the mean difficulty of the items of the four games is quite another matter. The Trios game may be the generally easier of the four, and its extreme scores are quite moderate. The Bindings game stands out by having quite extreme minimum and maximum values.

5.2.2 Item task variation

A different way of distinguishing between the items of a game, would be to investigate the functions or behaviors involved in responding to each item. Do all items of a game activate the same set of 'mechanisms' in producing the correct response? Or is it more appropriate to view the items as challenges to executive functions that are entirely or partly different? May tasks be classified into separate groups based on the functions needed for success? And should the necessary functions be viewed as largely independent, or as (partly) overlapping or interdependent?

Questions of this kind will also apply to the four games. While Table 1 implies different functions, much of the literature emphasize the functions' interdependence. The question thus is: Do the four games tap executive functions *in general*, or do they measure a set of different functions that are analytically and empirically *separable*? And if different processes provide a tenable perspective, can the Yellow/Red measure them?

5.2.2.1 The Cat/Dog game

This game may provide a convenient point of departure, as it implies a distinction between activation and inhibition. Among its 33 items, 16 have been labelled as 'consistent'. Here, the relatively simple correct response implies pressing the button on the *same side* of the screen as the stimulus picture. The remaining 17 items are 'inconsistent', requiring a button on the *other side* to be pressed. This response is more likely to involve inhibition of 'wrong' responses and may thus reflect a less simple process.

For a preliminary check on this idea, the mean score of the 16 'consistent' items on the original scale was compared to the mean of the 17 'inconsistent' ones. The mean for 'consistent' items was, 0.51, while the 'inconsistent' was only 0.42. The difference is statistically significant ($t = 5.490$; $df = 137$; $p < 0.001$). It may also be noted that the correlation between the two measures was 0.62.

By analyzing data from the binary scale, we find that the 'consistent' items also yield a higher mean number of correct responses (7.31) than the 'inconsistent' (6.06). The difference is also statistically significant ($t = 4.439$; $df = 138$; $p < 0.001$).

No corresponding difference occurs with the difficulty-weighted data, however. Mean scores from 'consistent' and 'inconsistent' items are also not significantly correlated to any of the eleven indices of the BRIEF inventory.

Nonetheless, the differences found in the original scores as well as the binary, 'simplified' ones suggest that the distinction between 'consistent' and 'inconsistent' items may be worth a further look. And the 0.62 correlation may be consistent with a hypothesis of partial independence between activation and inhibition. A more careful testing of different factor models, e.g., may thus prove interesting.

5.2.2.2 The trios game

This game includes 21 items and is intended to assess cognitive flexibility. For each item, the task consists of selecting the three items *that have something in common* out of four figures presented. No explicit explanation is presented for what 'something in common' means in practice. Items 1-5, however; include three figures that have *color* as their common attribute. In items 6-10, the

common property is *shape*, while *size* is the critical attribute in items 11-15. In the remaining items (16-21), the three types of differences are mixed, in an irregular sequence.

Again, items may imply different challenges. As already shown in paragraph 5.2.1, item TR11 may be regarded as a 'pivot' point in the Triads game. People with correct responses to this item do far better in the rest of the game – and in the game as a complete entity. The point is not the difficulty of the item, however, but the task it implies. Following items where *color* and *shape* are the salient distinctions, item TR11 unexpectedly introduces *size* as the critical stimulus difference. Clearly, most respondents do *not* understand that a new stimulus variable has been introduced.

When *size* is unexpectedly introduced as the salient stimulus attribute, and most of our young respondents fail to grasp the point. Of course, this does indeed constitute a test on cognitive flexibility. It should be recognized, however, that with our sample this step has proven disproportionally difficult. The transition from *color* to *shape* was easier by far.

The very concept of 'game' may contribute to this problem, since Yellow/Red respondents are invited to play a set of games. Playing a game implies adherence to a set of rules that is specific to the game. Rules may be explicit, and even formalized in a written document. They may also be implicit, however. If implicit, they are informally derived from what appears as 'normal' behavior in the game. And, rather likely, our young respondents early in the Triads game correctly observe that *color* and *shape* are the relevant stimulus features. A game rule is inferred from this fact and is then followed.

It may be argued, therefore, that they misunderstand what the game is all about. The alternative would be to openly present the entire game as a challenge to their ability to find *new criteria* for sorting stimulus figures, and then include a training/habituation phase with more than two salient attributes.

At any rate, obtaining more even steps within the game would be an improvement. Some adjustment to this end would therefore be advisable. The reframing of the entire game just mentioned may be one possibility.

Another option may be to strengthen the reference to '*something in common*' in the oral instructions, perhaps through a reminder that several types of difference or similarity may prove relevant. Yet another possibility is to make the size differences between the stimulus figures larger and more obvious. All options would require considerable thinking, planning and experimentation, however; and will best be handled by the experienced people at the *Centro UC Tecnologias de Inclusión* in Chile.

5.2.2.3 The Arrows game

This game includes 36 items, and its purpose is to measure inhibition. As shown already in figures 23-25, the items produce strikingly different numbers of correct responses. Items ARR01 through ARR16 appear as rather easy, with an obvious predominance of correct responses. As an opposite, ARR17 stands out by yielding very few such responses.

Rosas et al. (2020) report the expose time for Arrows items to be 1000 msec, with a 500 msec interval. A closer look at game, however, reveals that while the times for items ARR01 through ARR15 were obviously longer – more than 2000 msec. For items ARR16 through ARR36, however, the exposure times may be in agreement with the numbers supplied by Rosas et al. (2020). Responding correctly to the late items is thus more difficult than with the early ones. It may be argued, therefore, that early and late items should be viewed as different item tasks.

The recorded reaction time to the Arrows item also witness to the difference between early and late items. In figure 45 on the opposite page, reaction times to the first 15 items are shown to be

clearly above 1000 msec. Reaction times to the 21 later items, however, are all below this limit. Obviously, the two sets of items have been handled differently.

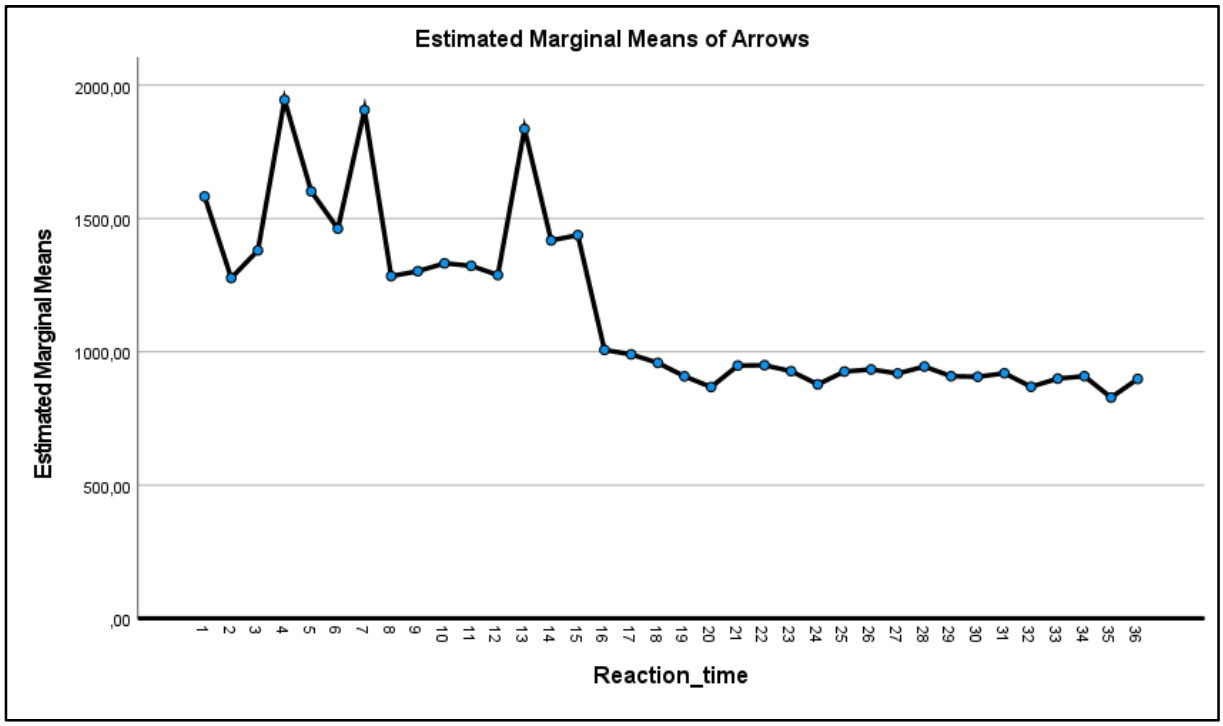

Figure 45: Reaction time to all 36 Arrow game items

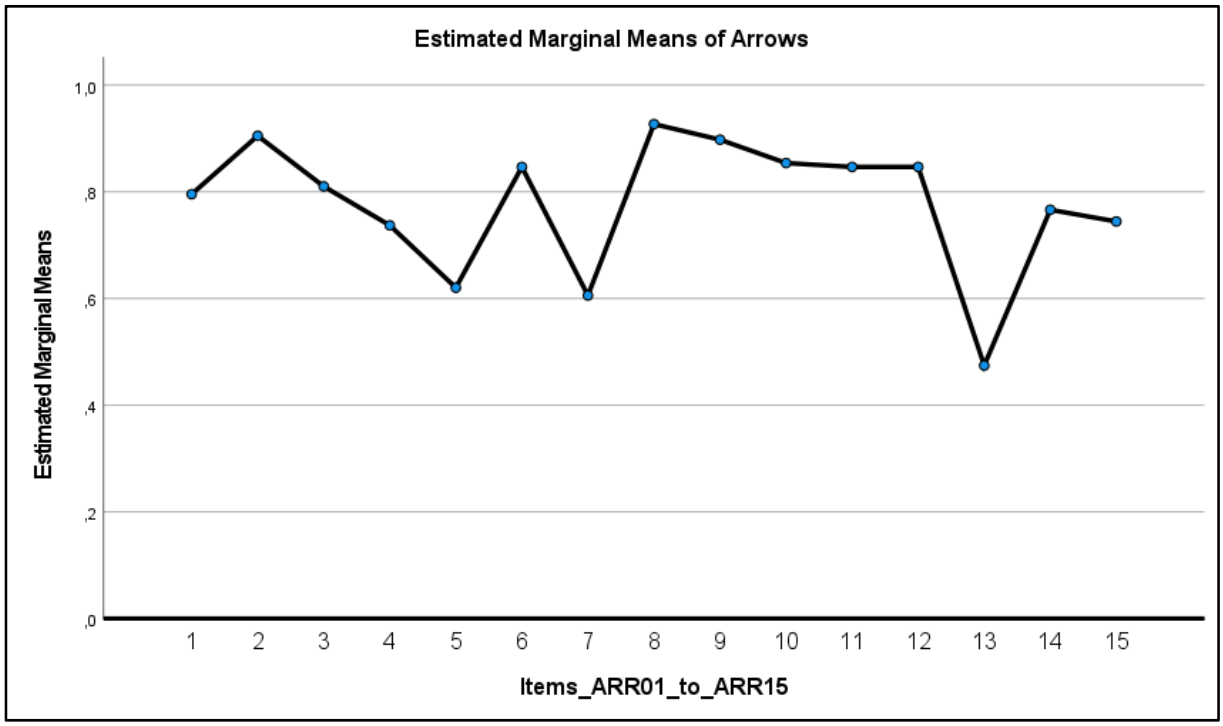

Figure 46: Fifteen early ARROW game items

And this difference of course affects the results of the 15 'early' (ARR01 – ARR15) items and the 21 'late' (ARR15 – ARR36) items. For the 15 early items, the grand mean of the summed Arrows game measure is 0.77, while the mean of the late 21 items is only 0.31. This difference is statistically significant ($t = 20.415$; $df = 138$; $p < 0.001$).

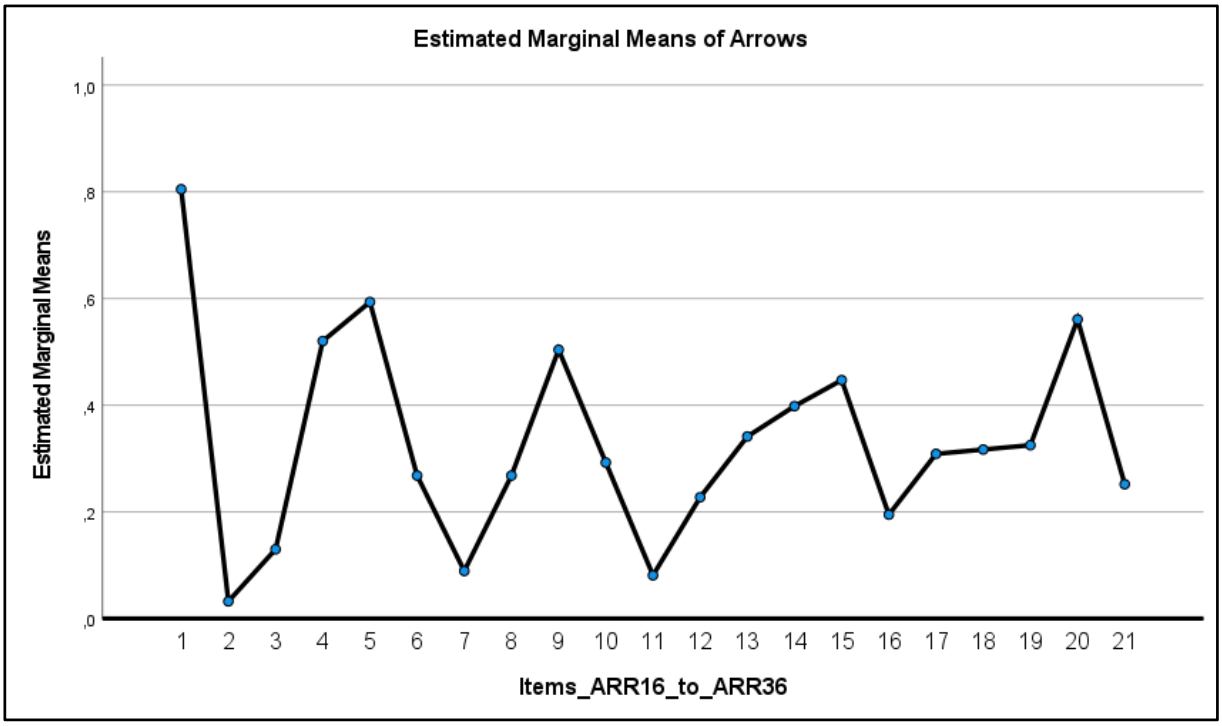

Figure 47: Twenty-one late ARROW game items

As figures 46 and 47 together show, the general levels of early and late Arrows items are indeed rather different. The figures also confirm the sizable item differences previously shown in figures 22 – 24. Consequently, both the early/late distinction and (additional) item difference may be considered examples of item task variation.

The 'standard' Yellow/Red coding procedure of summing all Arrows items into a common score, therefore, is running the risk of losing potentially interesting variance. One alternative approach could be to give more points to the 21 short-exposure items. Another option would be simply relegating the 15 long-exposure items to the status of non-counting 'practice items'. Also, as an opposite possibility, one could stick to long-exposure items only, disregarding the short-exposure ones as too difficult for our young sample.

It may be noted that treating the 15 initial items as constituting a single scale yields a Cronbach's *alpha* of 0.69. This equals the figure that table 2 showed for the pre-intervention Arrows scale. Viewing the 21 final items as a scale, however, gives an *alpha* only 0.55.

In table 14 on the following page, correlations between the two groups of Arrows items and the eleven BRIEF indices are shown. Six out of the eleven BRIEF indices correlate significantly with the mean of the 15 first (and long-exposure) Arrows items. With the 21 last (short-exposure) items, only two out of the eleven indices do.

A comparison to table 4 may also be in order. Here, the (original) 36-item Arrows scale is shown to be significantly correlated to five out of eleven BRIEF indices. This is rather similar to what table 14

shows for the 15 long-exposure items. For the 21 short-exposure ones, however, the picture of table 4 is rather different.

Table 14: Correlations between long- and short-exposure Arrows items and BRIEF indices

BRIEF Index		15 Longexp.	21 Shortexp
Inhibit N=135	Pearson r.	-0,09	-0,08
	Sig. (2-tailed)	0,28	0,38
Shift N=135	Pearson r.	-0,05	-0,13
	Sig. (2-tailed)	0,56	0,12
Emotional Control N=136	Pearson r.	-0,05	-0,07
	Sig. (2-tailed)	0,59	0,42
Initiate N=136	Pearson r.	-0,30	-0,16
	Sig. (2-tailed)	0,00	0,07
Working Memory N=134	Pearson r.	-0,35	-0,19
	Sig. (2-tailed)	0,00	0,03
Plan/Organize N=136	Pearson r.	-0,20	-0,02
	Sig. (2-tailed)	0,02	0,84
Organize Materials N=126	Pearson r.	-0,21	-0,11
	Sig. (2-tailed)	0,02	0,21
Monitor N=138	Pearson r.	-0,13	-0,07
	Sig. (2-tailed)	0,12	0,39
Behavior Regulation N=130	Pearson r.	-0,08	-0,10
	Sig. (2-tailed)	0,37	0,24
Metacognition N=120	Pearson r.	-0,30	-0,15
	Sig. (2-tailed)	0,00	0,11
Global executive N=112	Pearson r.	-0,21	-0,12
	Sig. (2-tailed)	0,03	0,20

Table 14 thus suggests that the long-exposure items may have a better validity than the short-exposure ones. In addition, the comparison with table 4 may imply that the validity of the complete Arrows scale receives more support from its 15 initial items than from the remaining 21. The better *alpha* of the long-exposure items scale may also favor this scale over the short-exposure one.

All in all, therefore, the question of dividing the original Arrows scale into 'early' and 'late' items appears to be warranted. Further analyses are certainly needed, however, before conclusions may be attempted.

Another difference between the Arrows items is the *direction* of the arrows. The 'model' arrow in the task is pointing either left, up, right, or down. Since the 'down' items are mainly tapping inhibition, they may be expected to differ from items with 'other' arrow directions.

A quick look back to tables 24 to 26 strongly supports this idea. The eight 'down' items are numbered 4, 7, 13, 16, 21, 25, 28, and 33. At all three time points, these items are the only ones yielding 'wrong' responses.

Table 15 on the next page also confirms that 'arrow direction' differences exist. They may not be exactly what one would expect, however. Only the Down items have negative scores (*wrong* responses), consistent with tables 24 – 26. Still, the mean scores of the Left and Right items are higher than those of *both* the Up and the Down items. Also, only the Down items have a mode of 1.00, showing that the most common response is 'correct'. In addition, Down items have a larger

standard deviation and as well as a more pronounced skewness than the others. Obviously, these items elicit different responses.

Table 15: Four arrow directions, basic statistics of all items

Statistics					
		Left	Up	Right	Down
		9 items	9 items	10 items	8 items
N	Valid	133	130	127	131
	Missing	17	20	23	19
Mean		0,53	0,44	0,60	0,46
Median		0,56	0,44	0,60	0,63
Mode		0,56	0,44	0,60	1,00
Std. Deviation		0,22	0,17	0,21	0,52
Skewness		0,03	-0,01	-0,38	-0,94
Std. Error of Skewness		0,21	0,21	0,21	0,21
Kurtosis		-0,29	0,63	-0,12	0,61
Std. Error of Kurtosis		0,42	0,42	0,43	0,42
Minimum		0,00	0,00	0,00	-1,25
Maximum		1,00	0,89	1,00	1,00

The differences are perhaps more easily visible in figure 48 below. The skewness towards score 1.00 (correct response) may be the most obvious difference between the Down (inhibition) items and the others. But the Down items are also the only ones yielding negative mean scores (wrong responses), resulting in a large range of mean scores.

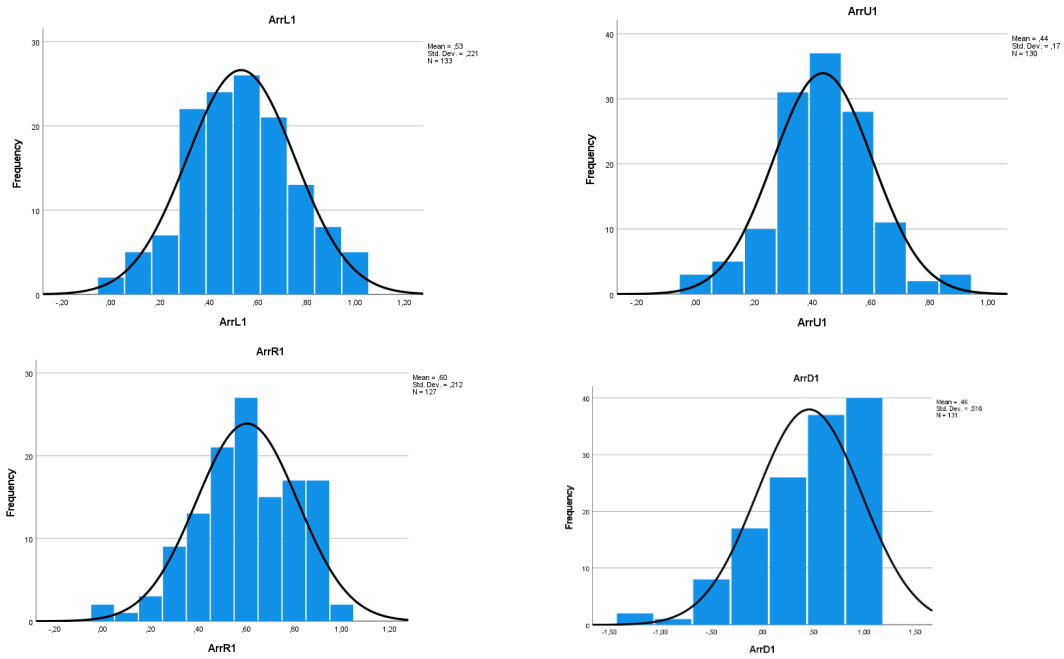

Figure 48: Distributions of left, up, right, and down arrow game items

An analysis of variance of the mean score values confirms that the mean scores of the four arrow directions are statistically different (MS = 0.603; F = 0.358; p < 0.001). Neither the intervention vs. control group difference nor the interaction was a significant effect, however.

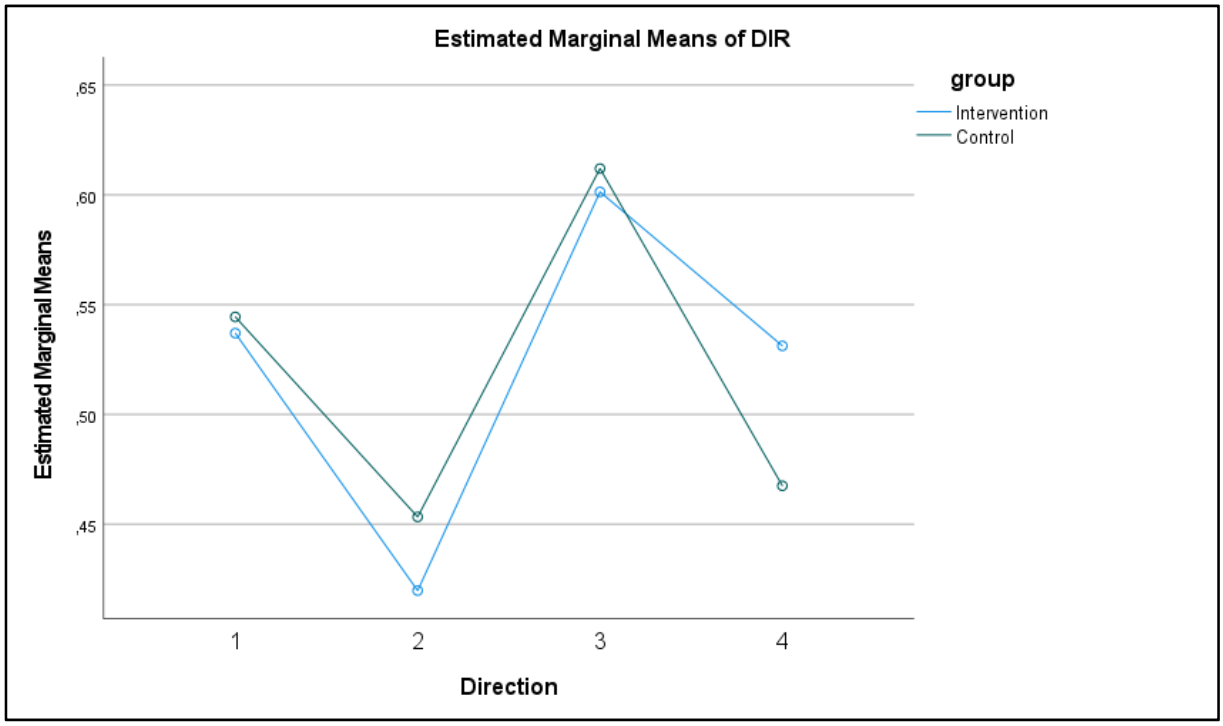

Figure 49: Means of intervention and control groups with items of four arrow directions

The intercorrelations of the four Arrow measures are far from perfect, as shown in table 16. The 'non-critical' Left, Up and Right item scales are consistently highly correlated. Again, however, the Down Arrow items prove different. Its correlation with the Left items scale is significant (r = 0.22 and p = 0.012), but the remaining two (Up and Right) items scales are not. Still, all four 'direction item' scales are highly correlated to the original Arrows summed scale.

Table 16: Four arrow directions, intercorrelations of independent scales in first round

Correlations

		ArrL1	ArrU1	ArrR1	ArrD1	ArrowsSum
ArrL1	Pearson r.	1	0,58	0,68	0,22	0,76
	Sig. (2-tailed)		0,000	0,000	0,012	0,000
ArrU1	Pearson r.		1	0,68	-0,030	0,61
	Sig. (2-tailed)			0,000	0,734	0,000
ArrR1	Pearson r.			1	0,050	0,72
	Sig. (2-tailed)				0,592	0,000
ArrD1	Pearson r.				1	0,71
	Sig. (2-tailed)					0,000

Clearly, the *down* items are different from others. This is no surprise, since they involve a different task. It means, however, that the scores from the original summed Arrows scale may not be the best approach to comprehend the trends in the data, since it lumps very different responses together. The interesting 'wrong' response, indicating a problem with inhibition, may be effectively 'hidden' in the composite score.

An alternative, therefore, would be to create a separate score for this particular response, keeping it apart from other interesting responses. Simply counting the number of 'wrong' responses for each respondent could be an option. As shown in figure 48, the range of this score proves to run from 0 to 8. The mode is 0, and most respondents have very low scores. It should be kept in mind, however, that there are only 8 Down items in the Arrows game. Since these items turn out to produce all the 'wrong' responses in this game, the generally low figures are not surprising.

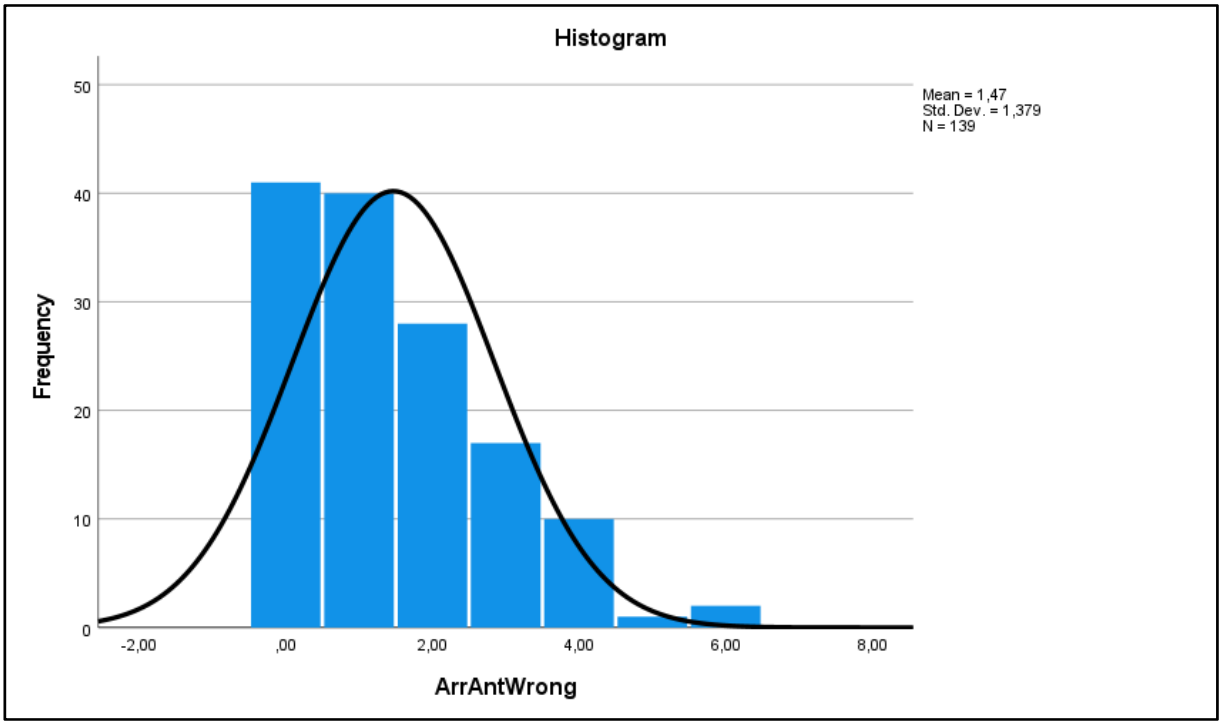

Figure 50: Number of 'wrong' responses in Arrows game.

Again, a comparison to the BRIEF indices may be useful. As shown in table 17 on the following page, the number of 'wrong' Arrow responses is significantly correlated to seven of the eleven BRIEF indices. In addition, correlations of three additional indices (Inhibition and Shift) are fairly close to significance. Only the Emotional Control index is clearly unrelated to the number of 'wrong' responses.

This compares rather favorably to the relations shown in tables 10 and 14. The number of 'wrong' responses in the first data round appears no less valid than the corresponding original Arrows scale or the long- and short-exposure items scales. Quite likely, the 'wrong' responses scale contributes strongly to the relevant variance of the original Arrows scale. But again, further analyses would be needed for more precise statements.

Table 17: Correlations between BRIEF indices and number of 'wrong' responses, first data round.

BRIEF Index		NoWrong1
Inhibit N=135	Pearson r.	0,14
	Sig. (2-tailed)	0,12
Shift N=135	Pearson r.	0,13
	Sig. (2-tailed)	0,14
Emotional Control N=136	Pearson r.	0,01
	Sig. (2-tailed)	0,89
Initiate N=136	Pearson r.	0,28
	Sig. (2-tailed)	0,00
Working Memory N=134	Pearson r.	0,37
	Sig. (2-tailed)	0,00
Plan/Organize N=136	Pearson r.	0,21
	Sig. (2-tailed)	0,02
Organize Materials N=126	Pearson r.	0,22
	Sig. (2-tailed)	0,02
Monitor N=138	Pearson r.	0,16
	Sig. (2-tailed)	0,05
Behavior Regulation N=130	Pearson r.	0,12
	Sig. (2-tailed)	0,17
Metacognition N=120	Pearson r.	0,33
	Sig. (2-tailed)	0,00
Global executive N=112	Pearson r.	0,27
	Sig. (2-tailed)	0,00

Compared to the original Arrows scale in the *third* data round, however, the 'wrong' responses scale of the first round falls short. As shown in table 12, the original Arrows scale in that round is significantly correlated to ten out of eleven BRIEF indices, including the Inhibition and Shift indices. The original Arrows scale thus appears to have some validity, even in the face of the reduced credibility of the BRIEF.

The 'wrong responses' scale, however, fares less well than the original Arrows scale in the third round. As shown in table 18 on the next page, the 'number of wrong' scale is significantly correlated to eight of the BRIEF scales in this data round. While this may be viewed as a slight improvement from table 17, it does not match the levels of table 12.

All inn all, therefore, the option of using the number of 'wrong' responses as a separate scale or measure does not look sufficiently promising to warrant further efforts within our present context.

Table 18: Correlations between BRIEF indices and number of 'wrong' responses, third data round.

BRIEF Index		NoWrong3
Inhibit N=97	Pearson r.	0,29
	Sig. (2-tailed)	0,00
Shift N=99	Pearson r.	0,19
	Sig. (2-tailed)	0,05
Emotional Control N=96	Pearson r.	0,14
	Sig. (2-tailed)	0,18
Initiate N=97	Pearson r.	0,20
	Sig. (2-tailed)	0,05
Working Memory N=99	Pearson r.	0,26
	Sig. (2-tailed)	0,01
Plan/Organize N=97	Pearson r.	0,19
	Sig. (2-tailed)	0,07
Organize Materials N=98	Pearson r.	0,08
	Sig. (2-tailed)	0,41
Monitor N=99	Pearson r.	0,31
	Sig. (2-tailed)	0,00
Behavior Regulation N=93	Pearson r.	0,25
	Sig. (2-tailed)	0,02
Metacognition N=92	Pearson r.	0,27
	Sig. (2-tailed)	0,01
Global executive N=88	Pearson r.	0,28
	Sig. (2-tailed)	0,01

Yet another difference between the 36 Arrow items may be briefly mentioned. A visual inspection of figures 24 – 26 show that the five short-exposure 'down' items (i. e. ARR16, ARR21, ARR25, ARR 28, and ARR33) are generally followed by a disproportionally high number of 'no responses'.

In our material, the 'wrong' code of course only occurs where no response should have been given. But why is no 'wrong' responses given to the following items? And why is their potential 'wrong' response apparently replaced by the 'no response'?

In part, this response pattern may reflect our young respondents' problem with this task. After the difficult inhibition-laden 'down task, inhibition apparently becomes dominant and commonly leads to 'no response'. Better explanations should probably be sought for this pattern. This would require more analyses, however, and will not be attempted in the present report.

5.2.2.4 The Bindings game

The task of this 27-item game is to remember which numbers are connected to specific figures. The procedure differs from that of the other games on two respects. The first difference is that the game is suspended after 3 consecutive errors. This means that after 3 incorrect responses, the subject automatically receives 'missing data' for all remaining items.

The second difference is that the number of figures in each item increases. In items 1 – 10, only two figures are to be matched with numbers. In items 11 – 22, three figures are presented, while items 23 – 27 contain four figures. The exposure time of the items also increases, from five to seven seconds. In spite of the increased exposure time, the items with three or four figures are clearly more difficult than the two-figure ones.

Figure 8 and figures 27 – 29 confirm that the 27 items are not equally difficult. In the first five items, responses are generally correct, while the next five also yield some 'wrong' responses. From item 11 on, the 'missing data' response quickly increases to become the most common one towards the end of the game. Quite likely, therefore, also the level of difficulty is a useful approach to the between-item differences of this game.

With our sample, the 27 items may be conveniently divided into four groups. In the first group (items 1-5), respondents generally give 'correct' responses. In the second (items 6-10), more errors appear. The third (Items 11-22) yield increasing numbers of missing data. The fourth item group (items 23-27) contains 'level 3' items; and produced no 'correct' responses in our young sample.

Naturally, the mean 'Bindings' score was different in the four item groups. It was close to 1 with the first item group, where most responses were 'correct'. It then gradually dwindled down to 0 in the final item group.

More important, however, is the number of missing data produced in the scoring of this scale. Before reaching the fourth item group, 137 respondents out of our sample of 139 had given 3 incorrect responses and was removed from the game. The final five items were thus not presented to them. Its mean of 0 thus is the score for the two respondents that 'survived' the items of Group 3. Consequently, no other statistics were produced for item group 4.

For item groups 1-3, however, the additional statistics are interesting. Median and mode both decrease from 1 through 0.6 to 0, while the dispersion (standard deviation) increases from 0.05 through 0.22 to 0.34. Corresponding changes are also seen with skewness and kurtosis. While the very narrow distribution of item group 1 is extremely skewed to the right; the distribution of group 3 is more moderate, and slightly skewed to the left.

Table 19: Basic statistics of four item groups in the Bindings game, first data round.

		Item Group 1	Item Group 2	Item Group 3	Item Group 4
N	Valid	139	139	132	2
	Missing	0	0	7	137
Mean		0,99	0,67	0,28	0,00
Median		1,00	0,60	0,00	0,00
Mode		1,00	0,60	0,00	0,00
Std. Deviation		0,05	0,22	0,34	0,00
Skewness		-8,24	-0,41	0,83	
Std. Error of Skewness		0,21	0,21	0,21	
Kurtosis		66,94	0,81	-0,34	
Std. Error of Kurtosis		0,41	0,41	0,42	

The second and third round of data produce very similar results. A quick look at figures 28-29 will also confirm this.

Clearly, also the Bindings game contains items of different types. By lumping them all together in one common scale, we therefore run the risk of losing potentially interesting information.

Computing separate scales for the four item groups seems not to be an option, however. The low dispersion in item group one means that its scale shows no difference between the respondents. Even worse, the fourth item group comes close to having no responses.

An attempt to validate the four scales with the 11 BRIEF indices was not successful. Correlations of Item group 4 had no variance and could not be computed. Of the 33 remaining (11 x 3) correlations tested, only one was statistically significant. The Second item group scale was related to the Working Memory index ($r = -0.18$; $p = 0.04$), while no other relationship was statistically significant.

This is exactly the problem that the 'simplified' or binary scale of paragraph 5.1 was intended to avoid. By simply recoding all responses into either 'correct' or 'other' responses, the complete set of respondents are at least retained in the sample.

As shown in table 20, the means resulting from the new algorithm decrease from the first through the last item group. The change patterns of median and mode is also the same as in table 19, only with item group 4 added. The new coding does little to increase the variance of the item groups 1 and 4, however; and appears to reduce the variance of item group 3.

Table 20: Basic statistics of four item groups on 'binary' scale of the Bindings game, first data round.

		Item Group 1	Item Group 2	Item Group 3	Item Group 4
N	Valid	139	139	139	139
	Missing	0	0	0	0
Mean		0,99	0,66	0,07	0,00
Median		1,00	0,60	0,00	0,00
Mode		1,00	0,60	0,00	0,00
Std. Deviation		0,05	0,22	0,12	0,00
Skewness		-8,24	-0,43	2,21	
Std. Error of Skewness		0,21	0,21	0,21	0,21
Kurtosis		66,94	0,82	5,85	
Std. Error of Kurtosis		0,41	0,41	0,41	0,41

The only interesting change from table 19 is the increased N. No other change in the main statistics is produced by the alternate coding algorithm. The correlations with the BRIEF indices have also been shown to be the same. The simplified or binary recoding thus offer no improvements to the four item group scales of the Bindings game.

6. SUMMING UP

The intention of this report has been to identify interesting and challenging questions for further work on the Yellow/Red test battery. A few options or alternative procedures have been sketched, but mainly to illustrate that different approaches may be viable. Consequently, these ideas should only by viewed as suggestions for possible modes or directions of thought, not as fully developed alternatives. We hope to raise interesting questions, not to have present their solution.

6.1 ORIGINAL SUMMED SCALES

The main 'outcome' variables of the Yellow/Red are the measurement scales from its four games. Some attention is due, therefore, to these scales. In old-fashioned psychometrics, reliability and validity are central in the evaluation of measurement scales (Cf., e.g., Christensen et al., 2011).

For two of the four Yellow/Red games, the measurement looks *reliable*. With our sample, the *Cronbach alphas* of the Cat/Dog scale and the Arrows scale are acceptable. The Trios scale is not quite up to these standards, however; and a large number of missing data prevents a simple computation of the *alpha* of the Bindings scale. This leaves room for improvement.

The *validity* of the four scales is also worth discussing. As shown in table 1, they are intended to measure partly different 'things'. *Cognitive flexibility*, *inhibition* and *working memory* (and their combination) may all be viewed as real-life functions, abstract phenomena, or theoretical concepts. At any rate, however, their measurements should be accurate: They should measure what they should and nothing else. Since the three concepts are known to be highly interrelated, this separation is a daunting task.

A common approach to assessing the validity of a test is to use an alternate, credible test for comparison. If the results of the two tests coincide, then the 'new' test is viewed as valid. In our material, the BRIEF inventory apparently offers independent measurements through indices that look partly similar to the three main concepts of cognitive flexibility, inhibition and working memory.

As shown in tables 4 – 5, comparing the BRIEF indices to the four Yellow/Red scales is not a clear success. In the two initial data rounds, the correlations between indices and scales are not generally impressive. Table 6, however, shows more promising results from the third data round. Here, most Cat/Dog and Arrows measurements are significantly correlated with the BRIEF indices. The fact that the third data round produces a different result may need an explanation. The decrease in the number of respondents in this round may be noticeable, but how it influences the correlations is not clear.

Viewing reliability and validity together, some differences between the four scales become obvious. The Cat/Dog and the Arrows scales both have acceptable reliability. In the third data round they also exhibit interesting validity properties. The Trios and the Bindings scale fare way worse. In our simple computations, neither reliability nor validity is convincingly demonstrated.

6.2 LONGITUDINAL CHANGES

For all four scales, scores are generally improving over time, as shown in figures 3, 5, 7, and 9. This also hold for the general, combined scale (Suma Z de Pruebas) of the entire Yellow/Red.

It is important not to misunderstand this trend, however. It does *not* show an effect of the intervention of the project. ANOVA shows no interaction affect; the increase in scores is found in the control group as well as in the intervention group. Learning or maturation effects are thus more likely than the project intervention to explain the observed response patterns.

The mean of the original scales is not the only interesting trend in the data, however. The distribution of response *types* also deserves some attention. As seen in figures 14 – 17, the percentage of *correct* responses generally increases over time. This does not come at the expense of *wrong* responses, however; their numbers do not change much. The corresponding change mainly occurs with the *no response* option.

This should perhaps invite to a closer look at the coding paradigms of the four scales. Here, numerical values are substituted for the different response categories, leaving the weighting of the categories implicit. Second, the different numbers are simply added together into what effectively is a composite score. The summed score thus contains information about quite different types of response. While this may be a practical and correct procedure for arriving at comparable numbers and trends, it may also serve to hide potentially important differences between respondents as well as between test items.

6.3 WITHIN-GAME ITEM DIFFERENCES

In the four games, between 27 and 36 different items or tasks are included. And generally, the items of a game yield very different responses. To some items, most respondents produce a 'correct' response. These items may thus be viewed as 'easy'. To other items, 'correct' responses are rare or non-existent, showing these items to be 'difficult'. In all games, most items fall in between these extremes. These items do produce some 'correct' responses, but not many; and must be classified as neither 'easy' nor 'difficult'.

As shown in chapter 4, these item differences are remarkably consistent across the three rounds of data collection. Items that are 'easy' at the initial point of time, generally stay so – and vice versa. This inter-item variance is lost, however, when the scores of all items are simply lumped together in a summed score. Utilizing this variance through an alternate coding paradigm may thus deserve consideration. Our modest attempts to approach this matter through simple recoding were not very successful, however.[5]

Quite likely, however, certain items may be more closely related to the purpose of the test than other items. The Arrows game, e.g., is intended to measure inhibition, and for 8 out of 36 items the 'correct' response is indeed to do nothing (not respond). And in the Cat/Dog game, the 'consistent' items may tap simpler processes (mainly action) than the 'inconsistent' (inhibition *and* action).

Differences of this kind may also inspire alternate coding schemes for the game scores. By weighting highly relevant items more than moderately relevant ones, the composite score may perhaps preserve more interesting variance than the original additive score. Our initial approaches

[5] As suggested in paragraph 5.2.1, Rasch scaling may be a more useful approach to this challenge.

to this, however, was clearly too simple. More sophisticated views, perhaps in the line of complex model testing, may have a better chance of making sense of the between-items variance concealed in the items' closeness or relevance to intended measurement.

6.4 CONCLUDING REMARKS

The basis of the present effort has been a data set that is limited in size, as well as in scope. The group of 149 informants is not large, and mainly supports simpler analyses. Consisting of only Norwegian school children from neighboring communities and born in two adjacent years, the sample also is rather homogeneous. Interesting variables and comparisons have thus not been available. Larger samples with more diversity would thus offer some advantages for further work on the Yellow/Red.

Hopefully, data from larger samples will also include more detailed and trustworthy information on respondents' executive functions than the BRIEF. In that case, also more reliable conclusions on the validity of the four test games could emerge from the comparisons between this information and data from the Yellow/Red scales.

With larger samples, also our ideas about alternate coding paradigms for the Yellow/Red could be put to a better test. The present attempts at recoding were not entirely successful, and neither prove nor disprove much. They may nonetheless point to alternate approaches to acquiring data from the Yellow/Red test battery, some of which deserve more sustained and qualified attention.

Our preferred focus would then be extracting more information from the diverse response categories than what is available through the summed score. In this context, deliberate experimentation with the weighting of the components may prove useful. Simplifying the scoring procedure would be another challenge, and not only by contrasting the 'correct' responses to all others. Perhaps a small number of 'cleaner' scores could be developed, each covering one specific facet of the original summed score.

7. REFERENCES

Allison, P. (2002). *Missing data*. Sage. https://doi.org/https://dx.doi.org/10.4135/9781412985079

Andersen, P. N., & Finbråten, H. S. (2020). Unsatisfactory psychometric properties of the Norwegian Behavior Rating Inventory of Executive Function Teacher Form - a Rasch Measurement Theory Validation. *Nevropsykologi*, *1*, 12-21.

Andersen, P. N., Klausen, M. E., & Skogli, E. W. (2019). Art of Learning -- An Art-Based Intervention Aimed at Improving Children's Executive Functions. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01769

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd. ed.). Routledge.

Christensen, L. B., Burke Johnson, R., & Turner, L. A. (2011). *Research Methods, Design, and Analysis (11th ed.)*. Pearson.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037-2078.

Garolera, M. (2019). *YellowRed International Results* [Internal report]. Escuela de Psicologia de la Pontificia Universidad Católica de Chile.

Gioia, G. A., Guy, S. C., Isquith, P. K., & Kenworthy, L. (2000). *Behavior Rating Inventory ofExecutive Function*. Psychological Assessment Resources.

Hundevadt, M. O., & Klausen, M. E. (2019). *Kan kunst være nøkkel for utvikling av eksekutive funksjoner hos barn? Avsluttende rapport for forskningspiloten "Kunsten å lære"*.

Håkansson, U., Andersen, P. N., & Kleiven, J. (In preparation, 2022). *Norsk versjon av "Yellow-Red": Et testprogram på Android nettbrett for eksekutive funksjoner hos barn* (Skriftserien, Issue.

Kleiven, J., Sandholt, L., & Andersen, P. N. (2022). Norsk versjon av "Yellow/Red": Et testprogram på Android nettbrett for eksekutive funksjoner hos barn. In *Skriftserien* (Vol. 18). Elverum: Innlandet University College.

Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children *British Journal of Developmental Psychology*, *21*, 59-80.

McAuley, T., Chen, S., Goos, L., Schachar, R., & Crosbie, J. (2010). Is the behavior rating inventory of executive function more strongly associated with measures of impairment or executive function? *Journal of the International Neuropsychological Society*, *16*(3), 495-505. https://doi.org/https://doi.org/10.1017/S1355617710000093

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. McGraw Hill.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, Design, and Analysis. An Integrated Approach*. Lawrence Erlbaum Associates.

Pino Munoz, M., & Aran Filipetti, V. (2019). Confirmatory Factor Analysis of the BRIEF-2 Parent and Teacher Form: Relationship to Performance-Based Measures of Executive Functions and Academic Acheivement. *Applied Neuropsychology: Child*, *10*(3), 219-233. https://doi.org/https://doi.org/10.1080/21622965.2019.1660984

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut

Rosas-Días, R., Espinoza, V., Santa-Cruz, C., & Martínez, C. (2022, Mai 2022). *The Yellow Red Test. Preliminary Results of the Chilean standardization process* [Power Point presentation].

Rosas, R., Espinoza, V., & Garolera, M. (2020). Evidencia intercultural de un test basado en Tablet para medir las functiones ejecutivas de niños entre 6 y 10 años: resultados preliminares. *Papeles de Investigatión (CEDETI, Escuela de Psicologia de la Pontifica Universidad Católica de Chile)*, *2020*(12).

Rosas, R., Espinoza, V., Garolera, M., & San-Martin, P. (2017). Executive Functions at the start of kindergarten: are they good predictors of academic performance at the end of year one? *Studies in Psychology (Estudias de Psicología)*, *38*(2), 451-472.

Toplak, M. E., Bucciarelli, S. M., Jain, U., & Tannock, R. (2009). Executive functions: Performance-based measures and the behavior rating of executive function (BRIEF) in adolescents with attention deficit/hyperactivity disorder (ADHD). *Child Neuropsychology*, *15*(1), 53-72. https://doi.org/ http://www.ncbi.nlm.nih.gov/pubmed/18608232

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. *Introduction to Rasch measurement*, *1*(1), 1-24.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Educational Measurement Solutions.

Høgskolen
i Innlandet

Hensikten med denne rapporten har vært å finne fram til interessante spørsmål for det videre arbeidet med testbatteriet Yellow/Red. Ganske enkle analyser har vært utført, delvis på grunn av datautvalgets begrensede størrelse (N=148).

Testskårene forbedres over tid, i likhet med andelen riktige responser. De ulike test-leddene eller oppgaven i hvert spill gir ganske ulike responser, og disse forskjellene er forholdsvis konstante over flere gjentatte målinger. De originale summerte skårene synes likevel å gi rom for forbedringer både i reliabilitet og validitet.

Noen alternative kodingsmuligheter blir derfor antydet, men blir hverken bekreftet eller avkreftet av våre forsøk på rekoding. Det er likevel mulig at noen av omkodingsidéene kan bli utprøvd på en bedre måte med et større utvalg.

The intention of this report has been to identify interesting questions for further work on the Yellow/Red test battery. Rather simple analyses have been carried out, partly due to the limited size of the sample (N=148).

The test scores consistently improve over time, as does the percentage of correct responses. Also, items within the same game elicit rather different responses, and item differences remain rather constant across replications. However, the original summed scales may have room for some improvement in both reliability and validity.

Consequently, alternate coding paradigms are suggested. They are neither proven nor disproven, however, by our initial recoding attempts. With larger samples, the recoding ideas may be put to a better test.