Jo Kleiven and Anita Kaderják

THE YELLOW-RED TEST IN THE "ART OF LEARNING" PROGRAM

Results from Hungarian and Norwegian art-supported learning in elementary schools

University of Inland Norway

Oppdragsrapport 4 - 2025

Jo Kleiven, University of Inland Norway.
Anita Kaderják, T-Tudok Knowledge Management end Educational Research Centre, Budapest.

Utgivelsessted: Elverum

Forfatterne er selv ansvarlig efor sine konklusjoner.
Innholdet gir derfor ikke nødvendigvis uttrykk for Universitetets syn.

I Universitetet i Innlandets oppdragsrapportserie publiseres både internt og eksternt inansierte FoU-arbeider.

Universitet i Innlandet - Oppdragsrapport nr. 4/2025

ISSN: 2704-2685

ISBN digital utgave: 978-82-8380-538-3

ISBN trykt utgave: 978-82-8380-537-6

Sammendrag

Rapporten analyserer resultatene fra Yellow-Red testen ved begynnelse og avslutning av Art of Learning-prosjektet i Norge og Ungarn. Hensikten med testen var å undersøke om deltakelse i prosjektet bedret de såkalte eksekutive funksjonene hos skolebarn.

Men etter korreksjon for aldersforskjeller får barn som har deltatt i Art of Learning-aktivitetene og en kontrollgruppe som ikke deltok, de samme resultater. Dette betyr trolig at testen ***enten:***
1. Ikke måler eksekutive funksjoner på en god og relevant måte ***eller***
2. Viser at deltakelsen i prosjektet ikke påvirker barnas eksekutive funksjoner.

Rapporten gjengir også noen deler av datamaterialet mer detaljert, for å være sikre på at deltester eller bestemte respondenter ikke gir andre resultater enn det samlede materialet. Trolig kan disse analysene også være relevante for videre diskusjoner om prosjektets praktiske prosedyrer og om evt. videreføring.

Emneord:

Art of Learning, executive functions, Yellow-Red test

Oppdragsgiver:

Innlandet fylkeskommune

Abstract

The present report analyzes the results from the Yellow-Red test conducted at the start of the Art of Learning project and again as a follow-up test nearly a year after the project's conclusion in Norway and Hungary. The intention of this testing was to see if participating in the project improved the so-called executive functions in school children.

After correction for age differences, however, children participating in the AoL project and a non-participating control group got the same results. This probably means that the test ***either***
1. Does not measure executive functions in a satisfactory and relevant manner ***or***
2. Indicates that participation in the project does not influence the children's executive functions.

The report also summarizes some parts of the data material in additional detail, mainly to check that subtests or respondent subgroups do not yield results that are different from those of the complete data matrix. These additional analyses may also be relevant to further discussions on the practical procedures of the project and on its future.

Keywords:

Art of Learning, executive functions, Yellow-Red test

Financed by:

Inland County Council, ICC

Preface

The present report has a rather limited focus. It mainly considers whether the Yellow-Red test indicates an improvement or not in the executive functions of school children after participating in the Art of Learning project. Consequently, only test data is used in this report.

The project also contains other kinds of data, however. Observational data have been collected and coded, e.g., the BRIEF inventory (Gioia et al., 2015; Hendrickson & McCrimmon, 2019) has been used, and careful records have been kept of procedures and behavior. These materials will not be utilized in the present report, due to its precise focus and the limited time available. This choice does not imply, of course, that this information is irrelevant to more general assessments of AoL impacts on children.

Moreover, the impact of the AoL project on executive functions is not the only important topic in this project. Another question is how the very concept of 'executive functions' should be comprehended and possibly measured. A third challenge may be how to best identify the value of art-based education; should, e.g., terms like creativity, inspiration, or motivation replace the current emphasis on executive functions? The project also may be an interesting starting point for the development of innovative pedagogical ideas and practical procedures. Could, e.g., its emphasis on non-verbal approaches be preferable for children who do not easily profit from the verbal learning techniques so prevalent in traditional schooling?

While questions of this kind should be recognized as important and challenging, they fall outside the scope of the present report. Our mandate has been to simply investigate whether the Art of Learning experiences yield improved measures of executive functions or not. We have added, however, several more detailed data analyses, mainly to make sure that parts of the material (respondent subgroups, subtests) would not yield other conclusions than the complete data matrix. Hopefully, the additional analyses will also give useful information about the practical procedures of the project. Discussions of alternate concepts and pedagogical innovations, however, must be left to future authors.

We wish to thank Maria O. Hundevadt for her helpful comments on a preliminary version of this report.

Jo Kleiven Anita Kaderjak

Content

1. Introduction

The *Art of Learning* is an Erasmus+ partnership project[3] involving Hungary, Norway and the United Kingdom, focusing on the influence of certain art-based school experiences on children's executive functions. The executive functions (Diamond, 2013) were assessed by administering the *Yellow-Red test* before and after the art experiences in the schools.

The *Yellow-Red test* is a set of performance games for Android tablets developed by Ricardo Rosas and his colleagues at CEDETI (Center for Development of Inclusive Technologies) at the Pontificia Universidad Católica de Chile (Rosas-Días et al., 2019; Rosas et al., 2022). Using this tool, parallel investigations of executive functions have been carried out in schools in Hungary and in Norway.

In advance, a pilot study had been carried out in Norway (Andersen et al., 2019; Hundevadt & Klausen, 2019). However, its *Yellow-Red* data failed to show any effect of the art experiences on the executive functions of the children (Kleiven et al., 2022).

In Hungary, some analyses of the main project data have already been published. Zágon and Németh (2022) give an interesting review of the art-related material and the activities employed, and Németh (2023) reports that preliminary findings support the belief that project activities favorably influence the children's "*…executive functions and creative habits…*". Based on a more detailed analysis of the Yellow-Red data, however, Kaderják (2024) is more cautious. She finds no significant difference between the children exposed to the *Art of Learning* experiences and other children.

Since similar methods have been used, the better part of the data from the Hungary and Norway should be comparable. The purpose of the present article, then, is to exploit the possible advantages of this comparison.

1.1 Research questions

A central assumption of the project is that children's level of executive functions may be measured by the Yellow-Red test. The level of executive functions is also expected to improve with age, as do most other skills.

The Art of Learning program (2021–2024) was evaluated at four time points: before the intervention (baseline assessment), after the first year, after the second year, and one year following the conclusion of the intervention. Although intermediate assessments were conducted, this report focuses on the baseline and final measurement points. This decision was based on two primary considerations: (1) no significant changes were observed in the interim assessments[4] and (2) executive functions develop gradually over time, meaning measurable changes were expected to emerge after a longer period following the intervention.

[3] The Norwegian intervention is a cooperation between Innlandet fylkeskommune (Inland County Council, ICC), Kulturtanken (Arts for Young Audiences) and School dept. in Lillehammer, Øyer, Tynset and Alvdal Municipalities. Financiers are EU (Erasmus+), but also Kulturtanken, ICC and Sparebankstiftelsen DnB (The DnB Bank Foundation).

[4] Based on Anita Kaderják's work (not published).

After the initial testing, only the 'Treatment' group is exposed to a series of *Art of Learning* experiences. A similar Control group does not receive this treatment.

Our hypotheses were:

1. In the Treatment group, executive functions (Yellow-Red scores) will improve more than in the Control group.

2. General measurements of executive functions will generally improve with age.

The data material, however, may contain more information than what is directly relevant for simply testing the two hypotheses. Different parts of the data may well suggest other results (and even other questions) than the complete material. The test scores and the numerous background variables should thus be examined for additional knowledge about informant differences, about the four subtests of the summed measure, and about strengths and weaknesses of the data gathering procedures.

Quite likely, supplementary analyses will also be useful to the evaluation of the present project, and to the discussions of possible continued efforts. Examples of relevant practical concerns may be:

1. Will the four subtests separately yield the same results as the summed scores?

2. Except for age, will other respondent characteristics influence test scores?

3. Measurements should be similar in the two countries

4. Measurements should not be affected by the size of schools

2. Method

2.1 Samples

In both countries, the sampling basis was schools that wished to participate in the *Art of Learning* project (AoL). Here, students experienced a series of "art pedagogy" activities (Hundevadt & Klausen, 2019; Zágon & Németh, 2022).[5] In addition, comparable schools were selected for use as a Control group. In these schools, the data gathering and testing was the same as in the AoL schools, but they participated only in ordinary school activities.

2.2 Procedure

The Yellow-Red test has been revised during the last couple of years (Rosas et al., 2022), and a couple of new subtests (new 'video games' of *Flies* and *The Farm*) have been added to the four original ones (*Cat-Dog*, *Arrows*, *Binding*, and *Triads*). To maintain compatibility with the data from the Norwegian pilot study, however, it was decided not to include the two new subtests in the present project.

The four subtests were used 4 times during the project. In this study, we used the two data points that were the furthest apart to assess potential change: the initial testing in Autumn 2021 (hereafter: *Initia*l) - before the AoL groups were subjected to art experiences - and the follow-up testing in Spring 2024 (hereafter: *Later*) – one year after the intervention finished. Comparable data was also gathered from the Control group at the same time, but without involving art experiences in those schools. Information on Country (Hungary or Norwegian), Age (5 to 9 years), and Gender (boy or girl) was also noted, as was a code for belonging to the AoL or the Control group.

After each testing session, data from all tablets were electronically transferred for initial processing at the CEDETI. Here, individual scores for the four subscales were computed and noted, as well as their sum and other central parameters. The data set would then be available for further processing in Norway and Hungary.

Due to a minor application error, participants were not prompted for their birth date. Instead, the CEDETI program instead defaulted to an assumed age of 20 years for all students in both Norway and Hungary. Normally, standardized scores accounting for students' age would have been provided for further analyses. Since this was not feasible in our case, we had to rely on raw test scores instead.

In initial measurement analyses, however, we assessed whether there were significant differences between results derived from standardized and raw test scores. No statistically significant discrepancies were then identified (calculations of Kaderják). This suggests that the lack of standardization in this case is unlikely to have substantially affected the findings.

Following the test authors' (Rosas-Días et al., 2019; Rosas et al., 2022) recommendations, a *summed score* for each respondent was also computed by adding together scores from all four subtests. In addition, classroom observation and interviews were used; gathering information on the thoughts and experiences of students, teachers, school leaders, artists and parents. In Hungary, even the BRIEF

[5] For more detail, cf. chapter 2.3.

inventory was added to evaluate "*…everyday behaviors associated with…*" executive functions (Hendrickson & McCrimmon, 2019).

The combined Yellow-Red data set for Hungary and Norway was generously provided on short notice from Chile by prof. Rosa.

2.3 The intervention

The "Art of Learning" (AoL) is an educational intervention designed to support the development of students' executive functions (EFs) through arts-based learning. Grounded in performative learning theory and arts-based pedagogical approaches (Hundevadt, 2022; Østern et al., 2019), the AoL integrates structured creative activities into the school curriculum to facilitate cognitive and reflective engagement.

A central component of the AoL intervention is the use of standardized lesson plans, which are pre-scripted and identical across participating schools. The intervention comprises 216 hours of structured activities delivered over a 48-week period. Each session includes a warm-up, a main activity, and a reflection segment, following a systematic approach informed by existing theories on EF development, performative learning, and arts-based education. The program aims to foster skills such as inhibitory control, working memory, and cognitive flexibility (Andersen et al., 2019).

A key feature of the AoL is its structured collaboration between educators and artists. To ensure consistency in implementation, all participating teachers and artists undergo standardized training.

To facilitate scalability and replication, the AoL intervention is supported by an online platform, "The Art of Learning Handbook[6]." This resource provides access to the full set of lesson plans and training materials, allowing schools and educators to implement the program independently. The availability of structured guidance ensures that the intervention can be consistently applied across different educational settings.

The development and implementation of the AoL intervention are examined in a master's thesis (Hundevadt, 2022). This research explores the underlying discourses and theoretical perspectives informing the program's design, offering insights into its pedagogical implications. The study provides an analysis of how arts-based learning contributes to the development of executive functions and educational outcomes.

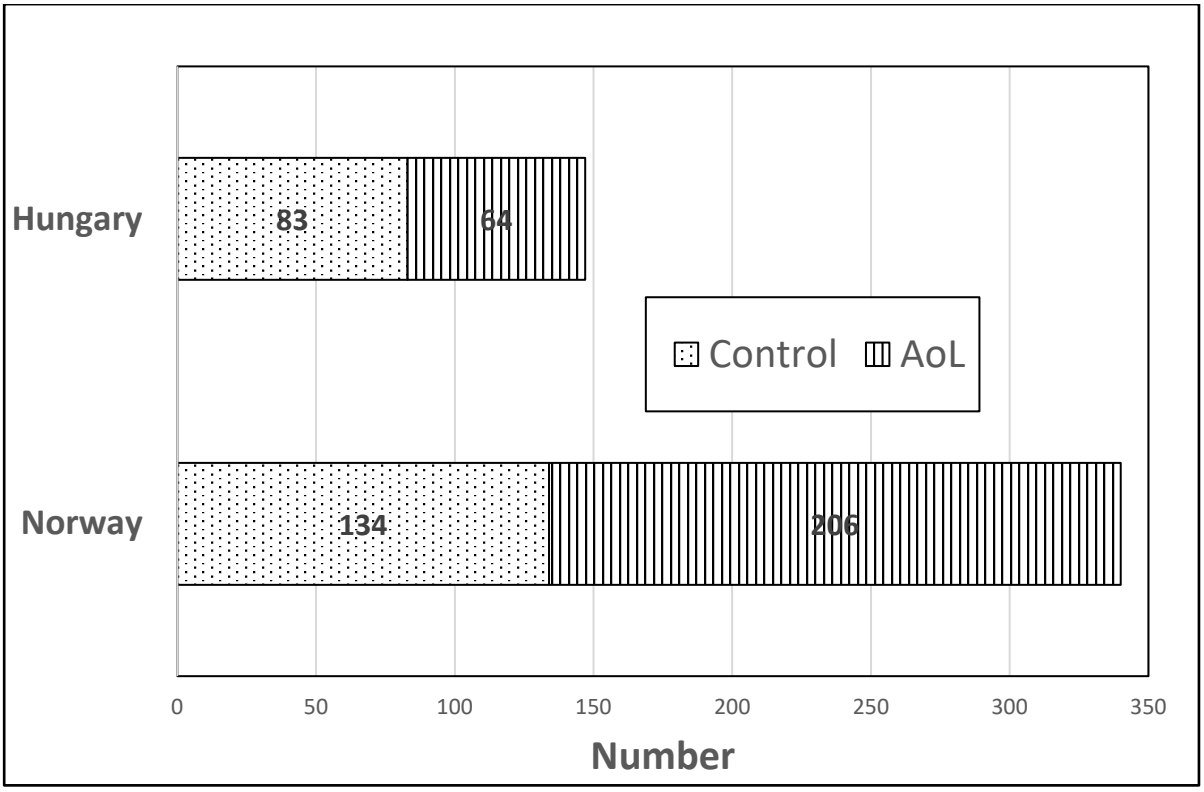
[6] https://artoflearning.t-tudok.hu/

3. Results

3.1 The samples

The samples turned out to be rather different in the two countries, as shown in Figure 1 below. Obviously, the two Norwegian groups are larger. But more importantly, the *Art of Learning* group is smaller than the Control group in Hungary, while the numbers for Norway show the opposite tendency. This difference is statistically significant (Chi-square = 12.078; df = 1; p < 0.001).

Figure 1: Art of Learning groups and Control groups in the two countries



The age distribution also exhibits national differences. This is shown in Figure 2. While the 6-, 7-, and 8-year-olds dominate the Hungarian sample, the Norwegian sample mainly include 6- and 7-year-old children. This difference of course is statistically significant (Chi-square = 93.591; df = 4; p < 0,001).

Figure 2: Age distribution in two countries

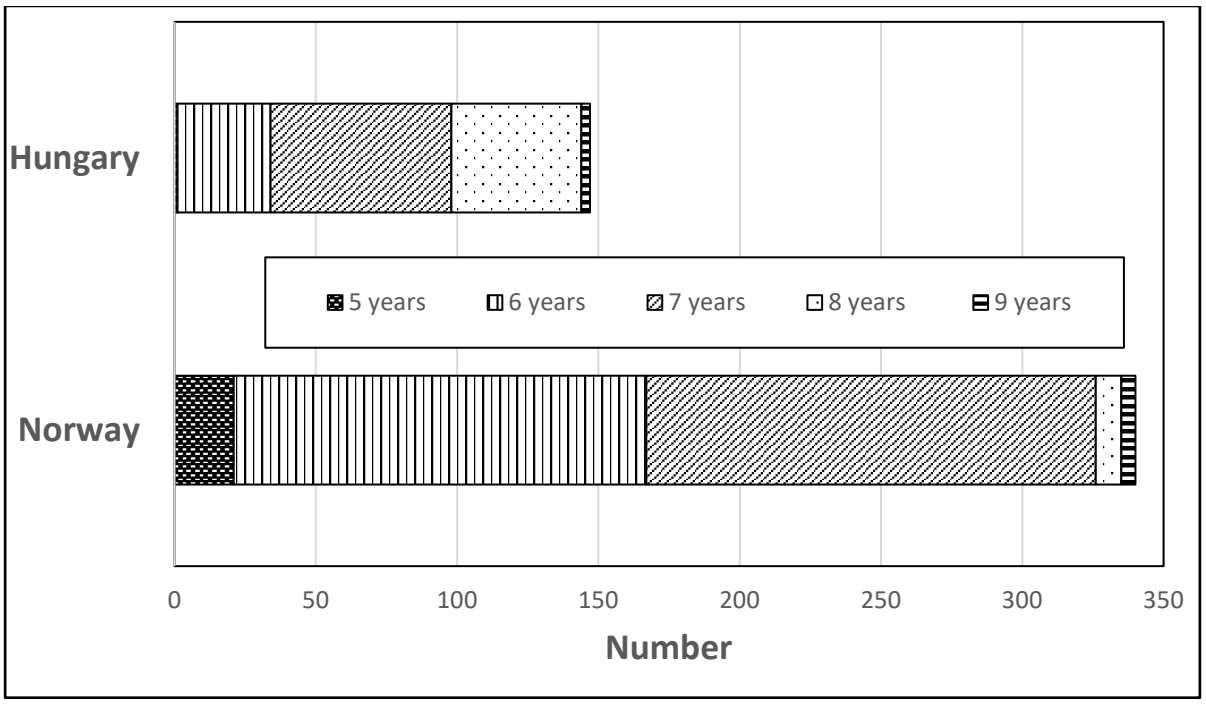
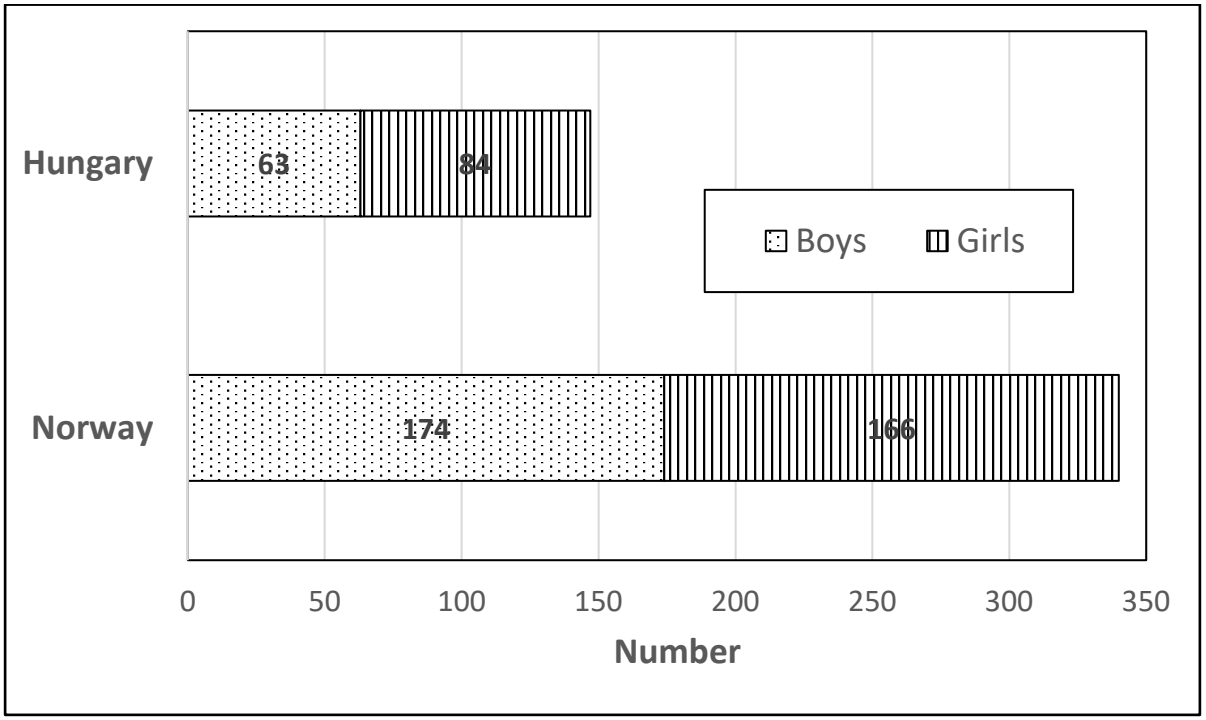


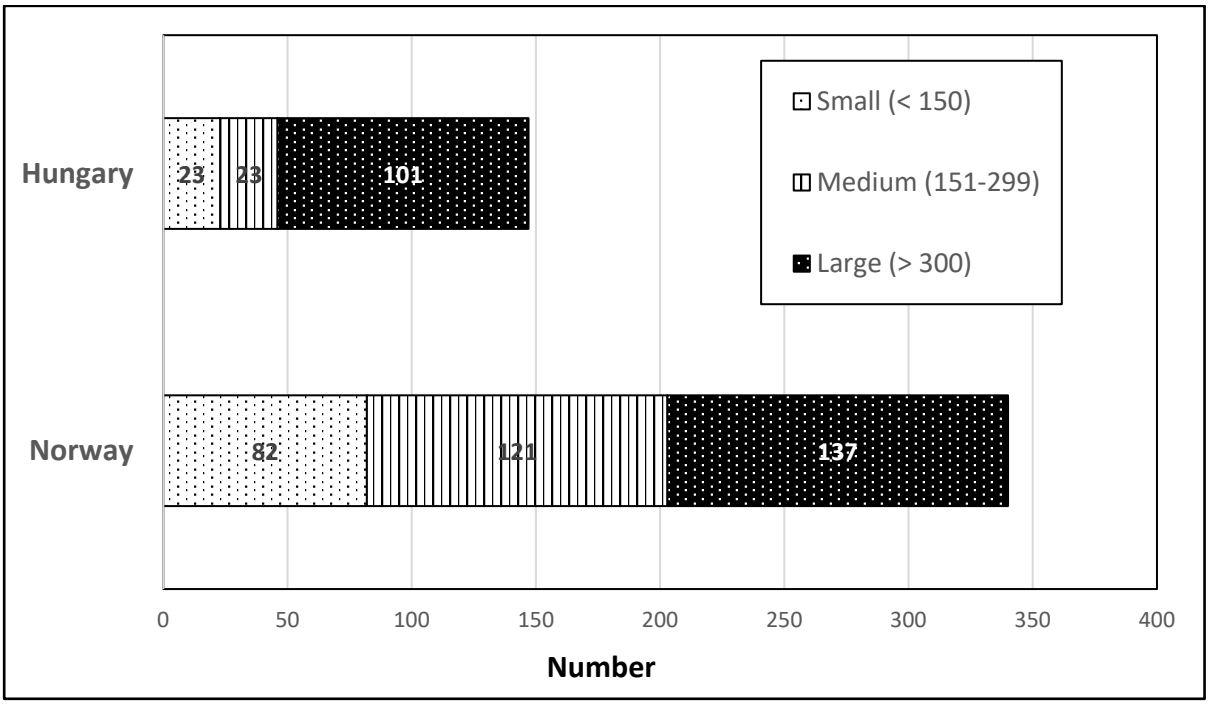
Figure 3: Gender distribution in two countries



A close look reveals that the Hungarian sample has a slight majority of girls, while the Norwegian has an even smaller majority of boys. This difference is not quite enough to reach statistical significance (Chi-square = 2.843; df = 1; p = 0.092).

In figure 4, the number of respondents attending schools of different sizes is shown for each country. Clearly, the proportion of children from large schools (69%) is higher in Hungary than in Norway (40%). And conversely, the proportion of children from small or medium size schools is higher in Norway (24 % and 36 %) than in Hungary (16% and 16%). The distribution is significantly different in the two countries (Chi-square = 34,173; df = 2; p < 0,001).

Figure 4: Number of respondents in small, medium, and large schools in the two countries



This makes it important to keep school size and country separate, and to control for school size when analyzing the effect of the 'country' variable. If not, misunderstandings are likely to occur.

More generally, the observed demographic differences between the Hungarian and Norwegian samples are not necessarily a problem. They should be kept in mind, however, when assessing the effect of the 'country' variable.

3.2 Summed Yellow-Red measurements

For each of the four subtests of the Yellow-Red, a score for each respondent is produced. In common use, these four individual scores (*Cat-Dog*, *Arrows*, *Binding*, and *Triads)* are summed into one common score for each respondent. This composite score is a simple sum, not involving any kind of weighting.

All measurements as well as their sum are recorded both *before* and *after* the AoL intervention. This provides a repeated-measure ANOVA design, where the repetitions (Initial vs. Later) constitute a within-subjects factor (Reps). Between-subjects factors include Country, Age, Gender, and Treatment (AoL vs. Control group).

3.2.1 Summed measures and all ANOVA factors together

In an exploratory initial step the effect of all relevant factors were assessed simultaneously. One was a within-subjects factor (Rep), covering Initial vs. Later test scores. Five were between-subject factors, covering respondent differences (Country, Gender, Age, Treatment, and School size). This roughly corresponds to using the ENTER procedure in a multiple regression, making sure that all available influences are assessed.

This preliminary analysis indicated that the Gender factor was far from statistically interesting (p = 0,977). Its numerous interaction effects also were not significant. In a second ANOVA, therefore, the Gender factor was removed.

The results of this second variance analysis are shown in table 1 on the following page. To enhance comprehension all statistically significant factors are marked with yellow.

Since it is closely related to the main purpose of the study, it should first be noted that neither the 'Treated' factor nor its interaction with Rep is statistically significant. The insignificant Treatment factor means that in general, the Yellow-Red scores (Initial as well as Later) from children who had shared the AoL experiences were no higher than the scores from the 'Control' children who had not. The *not significant* interaction effect is even more important, however. It means that the AoL treatment also has not affected the children's score *improvement* from the Initial to the Later testing – the improvement in the AoL Treatment group and the Control group is not different. These points will be further explained in paragraph 3.2.2.

The next two points are also central to the focus of the study. The Rep factor is clearly significant, meaning that the Initial Yellow-Red scores are different from the Later scores. In addition, the Age factor is significant, meaning respondents of different age generally have different scores on the Yellow-Red, i.e., both with Initial and Later scores. There also is an interaction between Rep and Age, however, meaning that the effect of Age is not the same on Initial and Later measures. Further discussion of this will follow shortly in paragraph 3.2.3.

The Rep factor also enters into a three-way interaction, involving both Country and Treatment factors. This rather complex result will be elaborated in paragraph 3.2.4. In addition, table 1 points out additional significant respondent differences. They relate less directly to the purpose of the study and will be treated in paragraph 3.2.5.

3.2.2 Summed measures, repetitions, and the AoL interventions

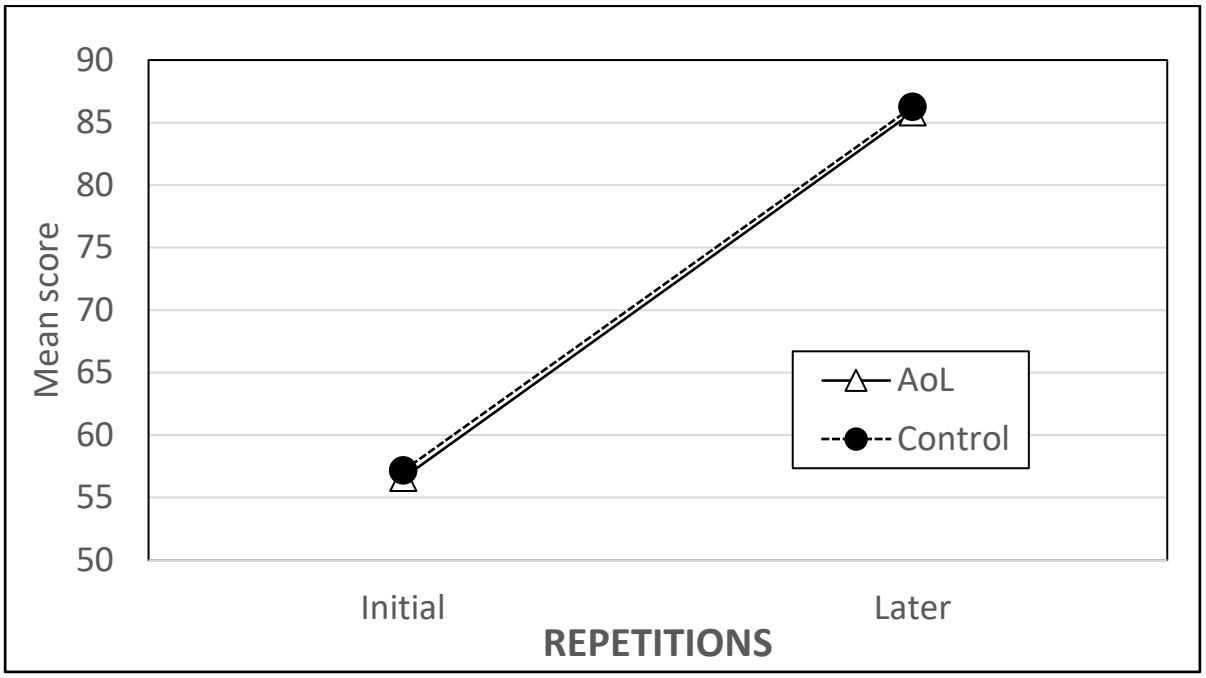
As shown in table 1, the *Repetitions* factor (i.e., *Initial* and *Later* scores) does make a difference to the Yellow-Red scores. The (*Treated*) difference between the AoL and the Control group is negligible, as is the interaction effect (*Repetitions/Treated)*.

A graph of the means involved may help to understand the meaning of these numbers. In Figure 5 (page 15), the means of the *Initial* scores are much lower than those of the *Later* scores. This holds for the *AoL* group as well as for the *Control* group. The graph also shows that there is virtually no difference between the *AoL* and the *Control* groups at any time. But more importantly, the small distance between the two groups is virtually unchanged from the *Initial* to the *Later* observations, so that the lines form close parallels. This indicates *no interaction effect*, i.e. the AoL "treatment" has made no difference to the summed score.

Table 1: ANOVA of summed scores before and after interventions in AoL and Control Groups

Within-Subjects Effects						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	
Rep	23555,087	1	23555,087	444,838	0,000	
Rep * Country	1,354	1	1,354	0,026	0,873	
Rep * Input_Age	2050,148	4	512,537	9,679	0,000	
Rep * Treated	34,239	1	34,239	0,647	0,422	
Rep * Size	123,922	2	61,961	1,170	0,311	
Rep * Country * Input_Age	131,306	3	43,769	0,827	0,480	
Rep * Country * Treated	207,238	1	207,238	3,914	0,049	
Rep * Country * Size	113,117	2	56,558	1,068	0,345	
Rep * Input_Age * Treated	60,666	3	20,222	0,382	0,766	
Rep * Input_Age * Size	355,495	8	44,437	0,839	0,568	
Rep * Treated * Size	14,146	2	7,073	0,134	0,875	
Rep * Country * Input_Age * Treated	184,267	2	92,134	1,740	0,177	
Rep * Country * Input_Age * Size	32,794	2	16,397	0,310	0,734	
Rep * Country * Treated * Size	32,739	1	32,739	0,618	0,432	
Rep * Input_Age * Treated * Size	63,052	5	12,610	0,238	0,946	
Rep*Country*Input_Age*Treated*Size	0,000	0				
Error(Rep)	23722,535	448	52,952			
Between-Subjects Effects						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	
Intercept	697347,750	1	697347,750	4449,893	0,000	
Country	506,530	1	506,530	3,232	0,073	
Input_Age	3122,293	4	780,573	4,981	0,001	
Treated	294,751	1	294,751	1,881	0,171	
Size	2060,524	2	1030,262	6,574	0,002	
Country * Input_Age	3184,244	3	1061,415	6,773	0,000	
Country * Treated	221,179	1	221,179	1,411	0,235	
Country * Size	969,773	2	484,887	3,094	0,046	
Input_Age * Treated	523,633	3	174,544	1,114	0,343	
Input_Age * Size	2745,593	8	343,199	2,190	0,027	
Treated * Size	216,674	2	108,337	0,691	0,501	
Country * Input_Age * Treated	528,740	2	264,370	1,687	0,186	
Country * Input_Age * Size	1703,283	2	851,641	5,434	0,005	
Country * Treated * Size	3182,977	1	3182,977	20,311	0,000	
Input_Age * Treated * Size	1097,732	5	219,546	1,401	0,223	
Country* Input_Age* Treated* Size	0,000	0				
Error	70206,579	448	156,711			

Figure 5: Means of summed scores before and after interventions in AoL and Control Groups



3.2.3 Summed measures at different ages and repetitions

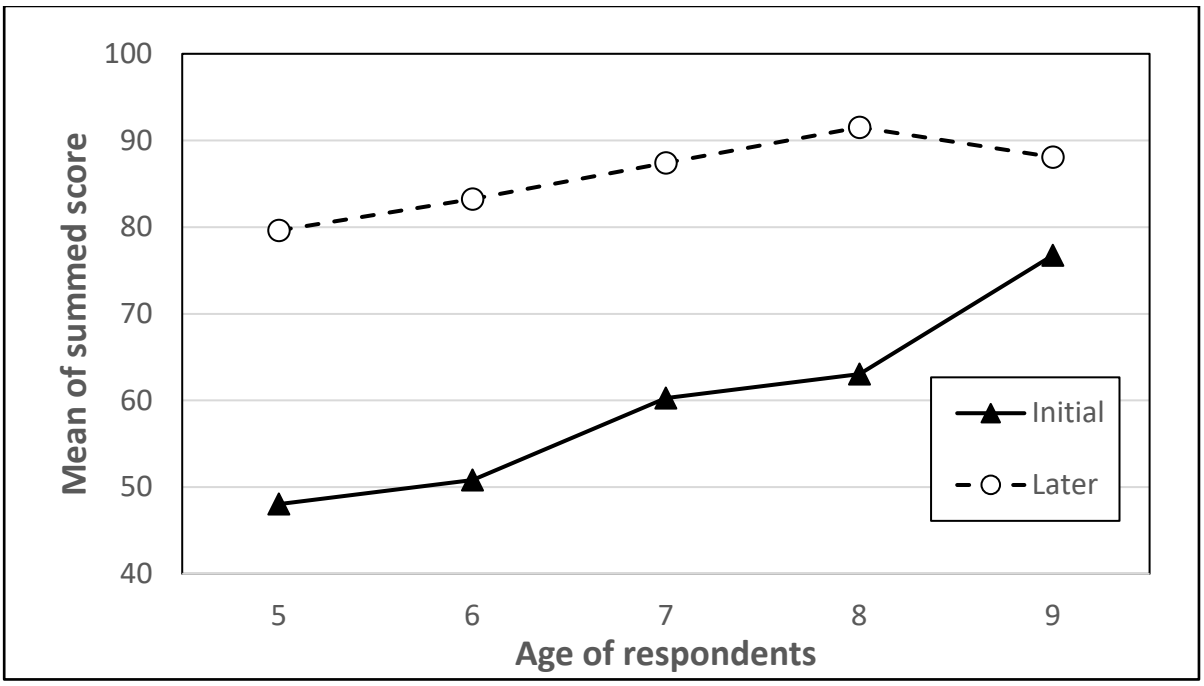
Another useful perspective may be to investigate the effect of the respondents' age on the summed scores. Table 1 indicated that the Rep factor and the Age factor were both significant, as was their interaction. As already shown, the initial scores are different from the later ones. But now also the between-subject effect of *age* is significant, as is the *reps* by *age* interaction effect. And again, plotting the means may help understanding the meaning of the numbers.

First, figure 6 on the following page confirms that the means of *Later* scores are higher than the *Initial* means. This holds for all five age groups, constituting the main effect of *repetitions*.

It is also clear that both means generally increase with age; both curves are slanted upwards. This is the main effect of *age*; children generally do better on the implied tests as they grow older.

But there also is a significant interaction effect, meaning that the effect of age is not quite uniform across all age steps. And the figure shows that at the age of 9, the common increase in *Later* responses does not occur. Instead, the *Later* means of the 9-year-olds drop down to the level of their 7-year counterparts. A ceiling effect may provide a likely explanation for this, however; the 9-year-olds perform at the end of the scale where further improvement is unlikely.

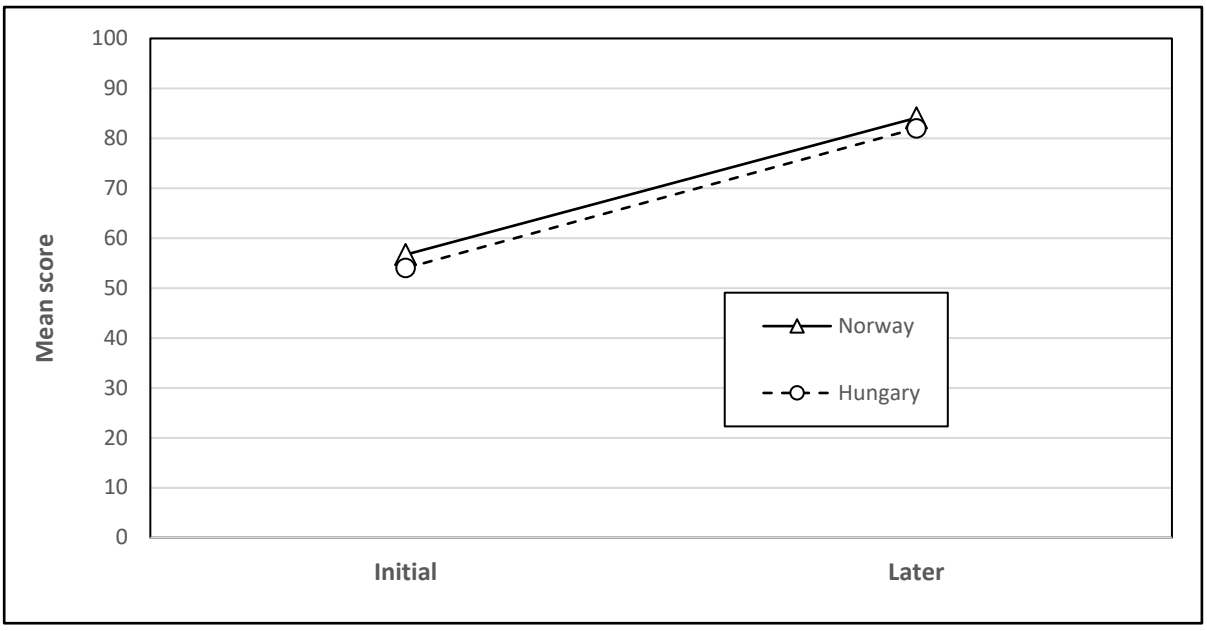
Figure 6: Means of summed scores before and after interventions in five age groups



3.2.4 The three-way Rep x Country x Treated interaction

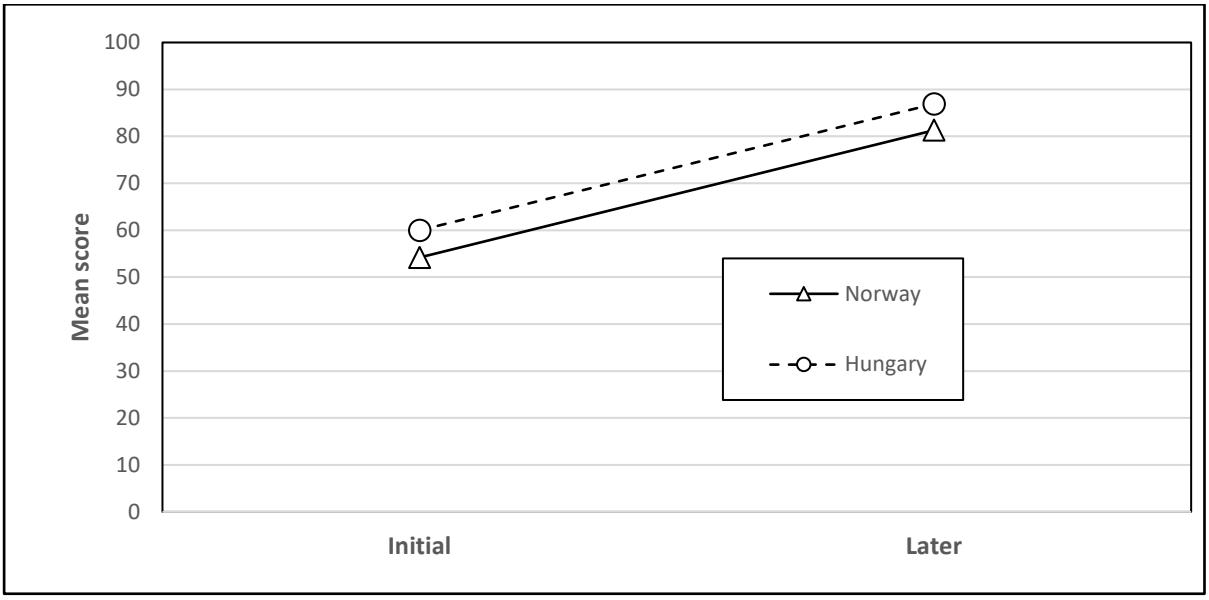
Two figures are required to understand this complex interaction. Figure 7 shows the mean scores for the respondents in the two countries that were not subject to the AoL experience, i.e. the Control groups. As expected, the Later scores are higher. The Norwegian scores are slightly higher than the Hungarian.

Figure 7: Means of Initial and Later summed scores in two countries, AoL Control group



The picture in Figure 8, however, is different. The Later scores are still higher than the Initial ones. But here the Hungarian scores are higher than the Norwegian – quite the opposite of the relationship shown in the previous figure. A way of describing the interaction effect, therefore, would be to say that the relationship between the Country and the Repeat variables depends on whether the AoL Treatment is given or not. For respondents getting this treatment, the Hungarians obtain better scores than the Norwegians. For respondents not subject to the AoL experiences, the Norwegians score higher.

Figure 8: Means of Initial and Later summed scores in two countries, AoL Treatment group



The three-way interaction also is only just significant (p = 0,049) and should probably be interpreted with some caution. It should also be noted that having many factors in the analysis implies a limited number of observations in some of the cells of classification. This may yield estimates that are not very robust, where small changes in the data may change results and conclusions.

This specific interaction effect is also not very relevant to the main findings on Repetitions, Treatments, and Age. All in all, therefore, it may not merit further attention.

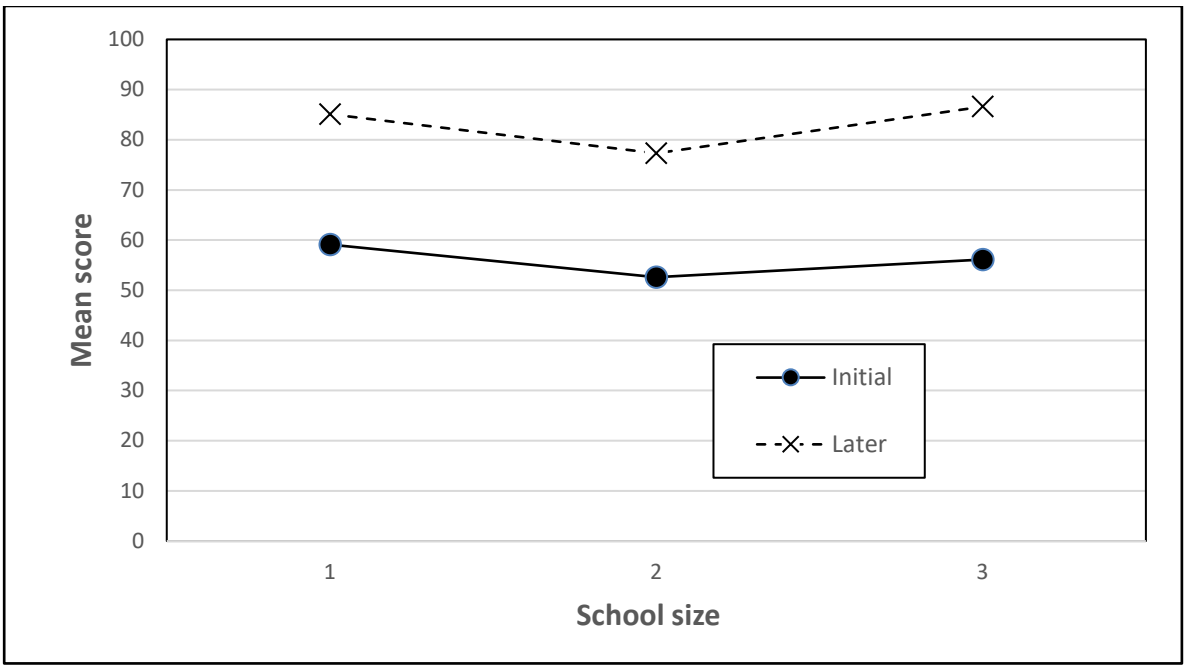
3.2.5 Summed measures and additonal respondent differences

Several respondent characteristics are shown to affect the summed Yellow-Red measurements, as shown in table 1. Respondents' age has already been mentioned (paragraph 3.2.3), but also School Size, Country, and Treatment prove to be interesting.

3.2.5.1 School size

The ANOVA in table 1 indicated that school size was a significant general influence in the Yellow-Red scores, i.e. both the Initial and the Later scores. Figure 9 on the following page shows the direction of these differences. The relationship here is not simple or linear; respondents from intermediate size schools receive lower scores than do respondents from both smaller and larger schools. This holds for Initial as well as Later scores. And throughout, the Later scores are of course higher than the Initial ones.

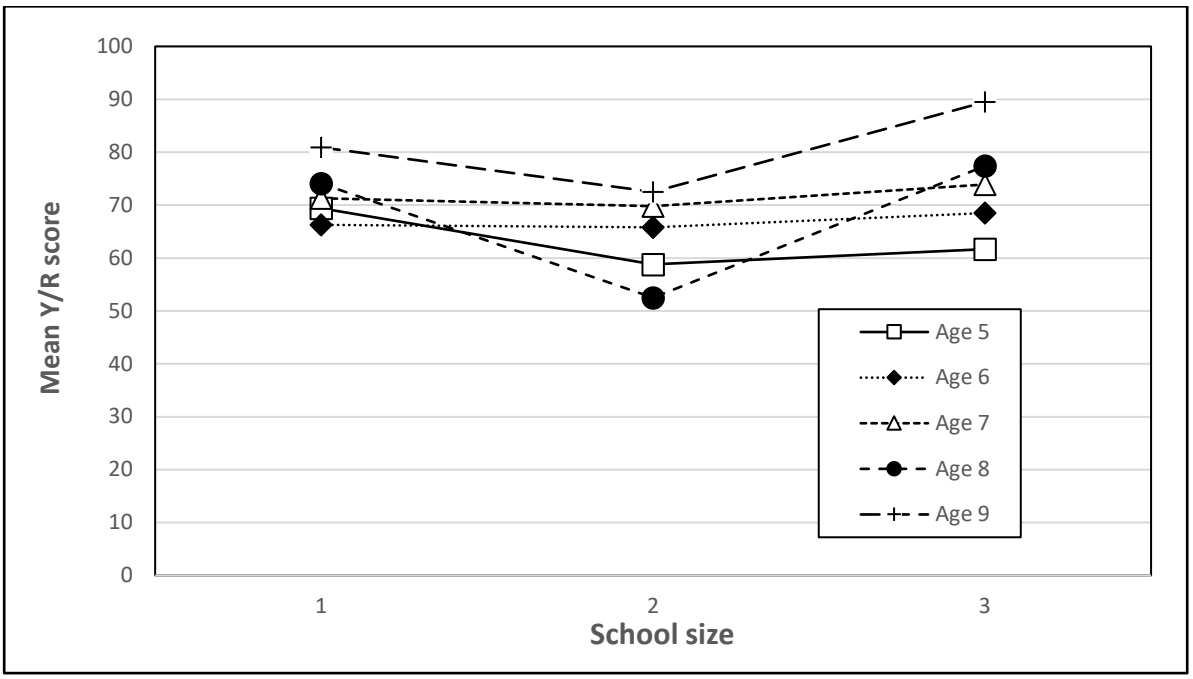
Figure 9: Summed score means from respondents in small, intermediate, and large schools



The figure as such may be clear and readable. This may be insufficient, however, for explaining or understanding it. Why does the intermediate schools produce lower Yellow-Red scores than smaller and larger schools?

School size also enters into a significant interaction effect with the Age variable. Again, a figure may be useful to comprehend the complex relationships involved. Generally, age groups follow the pattern that was shown in figure 9. Ages 8 and 9, however, have a much larger drop at the intermediate school size than the others. This appears to be the interaction effect.

Figure 10: Mean of two summed scores from respondents in small, intermediate, and large schools

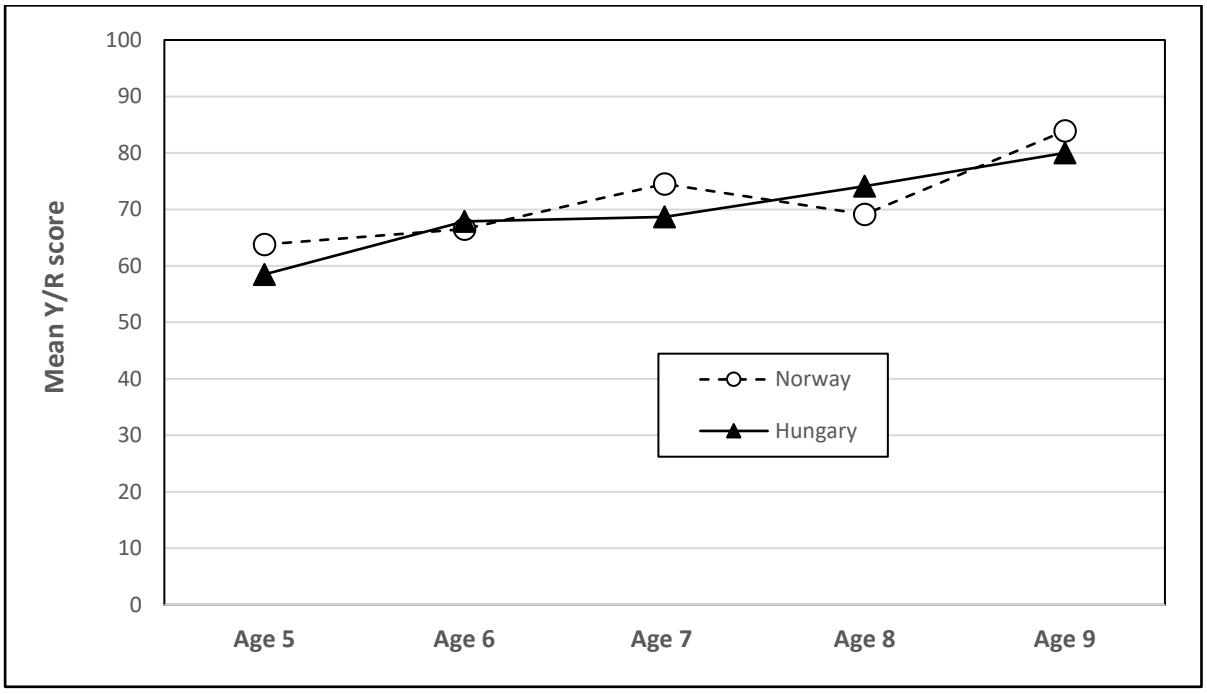


A closer look at the data, however, shows that there is only one respondent of Age 8 and one of Age 9 in the intermediate schools. This is unfortunate, since the very limited number of observations may be subject to random variations. No confidence can thus be attached to this specific outcome, and the interaction effect will not be further discussed.

3.2.5.2 Country

While Country in itself is a not quite significant predictor variable in table 1; it turns out to be involved in several interaction effects. The strongest of these is the Country x Age interaction ($p < 0{,}001$). The relationships involved are displayed in figure 11.

Figure 11: Mean of two summed scores from respondents by Age and Country

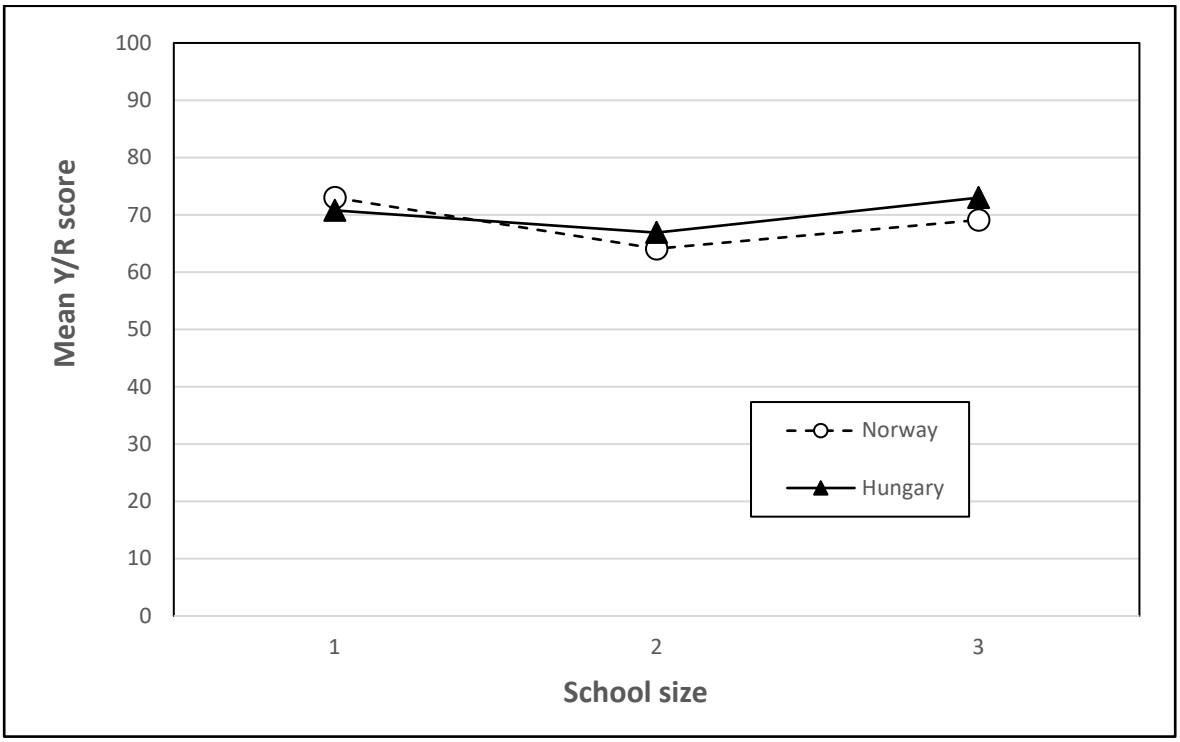


An apparently reasonable conclusion may be that the relationship between the two countries is dependent on the age of the respondents. For some age groups only, the Norwegian sample has the higher scores. For other age groups the opposite trend is apparent.

It should again be noted, however, that the number of observations is very low in certain cases. There is, e.g., only one Hungarian respondent in the 5-year group; and only three Hungarian 9-year olds. Some caution is thus in order for explanations of this interaction effect; even small random shifts in the data may change the apparent patterns.

The Country factor also (just) significantly interacts with School size ($p = 0{,}046$). Figure 12 on the following page may be useful for understanding the nature of this interaction. Firstly, the intermediate schools yield lower scores than others. Secondly, the lines of the two countries cross; while Norway has the higher scores in small schools, Hungary does so in the intermediate and large ones. This is the interaction effect: the difference between schools of different size is not the same in the two countries.

Figure 12: Mean of two summed scores from respondents by School size and Country



An additional complication, however, is that the Country factor also is involved in two significant three-way interactions. The *Country x Age x Size* interaction is statistically significant ($p = 0{,}005$), and so is *Country x Treated x Size* ($p < 0{,}001$).

Making sense of these interactions, however, requires breaking the data down into smaller groups. But in certain cases, there are no observations in these groups. In the Norwegian sample, e.g., there is no 9-year-old from the intermediate and large schools. And in Hungary the small and intermediate schools have no 5-year-olds. The means of these 'cells' may of course not be computed, and the significant Country x Age x Size interaction is in danger of being a partial artifact.

The Country x Treated x Size interaction is subject to a similar problem. In the Hungarian sample, there is no observations from the Control (No AoL treatment) group in the intermediate schools. Consequently, the mean of that 'cell' of observations cannot be computed. For the two three-way interactions, therefore, explanations would be based on incomplete data. A more prudent conclusion would be that more data is needed to establish more robust and trustworthy facts. The two interactions will thus not be further discussed in the present report.

3.2.6 Important influences on the Yellow-Red scores

A first conclusion may be rather straightforward: The summed scores generally improve from the initial repetition to the later. This improvement, however, appears not to be linked to the Art of Learning experiences; being about the same in the AoL and the Control group of students.

A second conclusion may be that the score improvement may be attributed to the increased age of respondents, since the respondents' age is significantly related to the summed scores. It should be borne in mind, however, that other variables are closely related to age, and thus may provide

alternative explanations. Previous experience with the test, e.g., may have yielded a 'training effect' that contributes to the improvement in summed scores (cf. paragraph 4.1.2).

In addition, School size may play a part here. Its effect is not linear, however; intermediate schools yield lower scores than small and large ones. Size also interacts with Country; its effect on executive processes is not quite the same in the two countries. Understanding this variable, therefore, is not straightforward (cf. paragraph 4.1.2).

The Country variable is also close to having a statistically significant effect on the summed scores. It may thus deserve consideration in more complex analyses. Factors Gender and Treatment appear not to influence the summed test scores very much.

3.3 Subtests of the Yellow-Red procedure

The results from the summed measures may seem discouraging, indicating no general effect of the AoL interventions. According to its authors (Rosas et al., 2022; Santa-Cruz & Rosas, 2017), however, the four tests are intended to measure partly different types of executive functions. It is tempting, therefore, to look for alternate ways of using the data from this test battery. A likely first step, then, would be to see if any of the four subtests (*Cat-Dog*, *Arrows*, *Binding*, and *Triads)* show other trends or yield different insights.

3.3.1 All four subtests in ANOVA of Initial and Later observations

In table 2, the four subtests are treated as repeated measurements in an ANOVA. A second repeated measurement is the Initial vs. the Later observations.

Naturally, the Rep factor is significant; merely confirming what has repeatedly been shown in the preceding ANOVA series. Later scores are higher than the Initial ones.

In addition, of course, the four subtests have significantly different means. This is no interesting surprise, since they have not been transformed to fit a common scale.

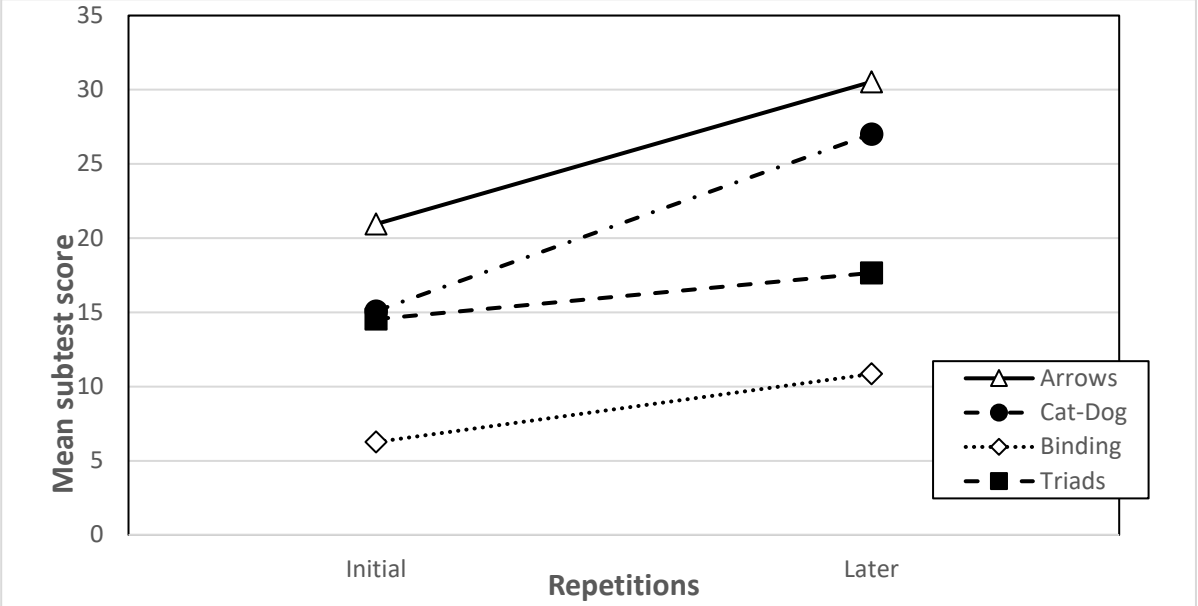
More important, however, is the significant interaction effect between the Rep and the Test factors. And again, a figure (Fig. 13 on the following page) will help to clarify the meaning of this effect.

The figure shows that the increase in score means is clearly stronger with the Arrows and the Cat-Dog subscales than with the Binding and Triads subscales. In other words, Arrows and Cat-Dog subscales contribute more to the previously observed increase in summed scores than do the Binding and the Triads ones. This may suggest that the intervention does not influence the four subscores equally. Simply summing the test scores, then, may hide interesting differences between the four games. It may thus be prudent to view the data from each game separately.

Table 2: ANOVA of four subtest scores before and after interventions

Within-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Reps	51621,957	1	51621,957	3548,190	0,000
Error(Reps)	7027,077	483	14,549		
Tests	156777,962	3	52259,321	2931,293	0,000
Error(Tests)	25832,889	1449	17,828		
Reps * Tests	12460,130	3	4153,377	359,230	0,000
Error(Reps*Tests)	16753,171	1449	11,562		
Between-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1235445,523	1	1235445,523	24003,649	0,000
Error	24859,561	483	51,469		

Figure 13: Means of four subtest scores before and after interventions



3.3.2 The Cat-dog task

This game is designed to measure inhibition; the ability to control impulses and resist distractions. The central question now is whether this score is subject to the same influences as the summed score or not. For direct comparisons, comparable ANOVA results should be an advantage.

3.3.2.1 ANOVA

Like in paragraph 3.2.1, an exploratory initial step will be to analyze all relevant factors simultaneously. Country turned out to be the least interesting factor here and was excluded from the main ANOVA. The results of the (simpler) variance analysis are shown in table 3.

Table 3: ANOVA of the Cat-dog subtest scores before and after interventions

Tests of Within-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Rep	5446,611	1	5446,611	304,922	0,000
Rep * Input_Age	424,428	4	106,107	5,940	0,000
Rep * Treated	3,952	1	3,952	0,221	0,638
Rep * Gender	4,802	1	4,802	0,269	0,604
Rep * Size	23,992	2	11,996	0,672	0,511
Rep * Input_Age * Treated	6,387	4	1,597	0,089	0,986
Rep * Input_Age * Gender	27,346	4	6,837	0,383	0,821
Rep * Input_Age * Size	163,410	8	20,426	1,144	0,333
Rep * Treated * Gender	1,976	1	1,976	0,111	0,740
Rep * Treated * Size	57,550	2	28,775	1,611	0,201
Rep * Gender * Size	25,905	2	12,952	0,725	0,485
Rep * Input_Age * Treated * Gender	2,429	2	1,215	0,068	0,934
Rep * Input_Age * Treated * Size	121,602	4	30,400	1,702	0,148
Rep * Input_Age * Gender * Size	25,453	3	8,484	0,475	0,700
Rep * Treated * Gender * Size	15,595	2	7,797	0,437	0,647
Rep* Input_Age* Treated* Gender* Size	86,415	3	28,805	1,613	0,186
Error(Rep)	7877,274	441	17,862		
Tests of Between-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	75332,038	1	75332,038	2102,989	0,000
Input_Age	2438,265	4	609,566	17,017	0,000
Treated	9,753	1	9,753	0,272	0,602
Gender	147,086	1	147,086	4,106	0,043
Size	299,062	2	149,531	4,174	0,016
Input_Age * Treated	37,727	4	9,432	0,263	0,901
Input_Age * Gender	221,687	4	55,422	1,547	0,188
Input_Age * Size	552,865	8	69,108	1,929	0,054
Treated * Gender	16,776	1	16,776	0,468	0,494
Treated * Size	89,637	2	44,818	1,251	0,287
Gender * Size	78,669	2	39,335	1,098	0,334
Input_Age * Treated * Gender	60,702	2	30,351	0,847	0,429
Input_Age * Treated * Size	135,042	4	33,761	0,942	0,439
Input_Age * Gender * Size	209,415	3	69,805	1,949	0,121
Treated * Gender * Size	6,249	2	3,125	0,087	0,916
Input_Age * Treated * Gender * Size	160,438	3	53,479	1,493	0,216
Error	15797,245	441	35,821		

As we can see, 25 out of 30 effects are not statistically significant. The only interesting effects are

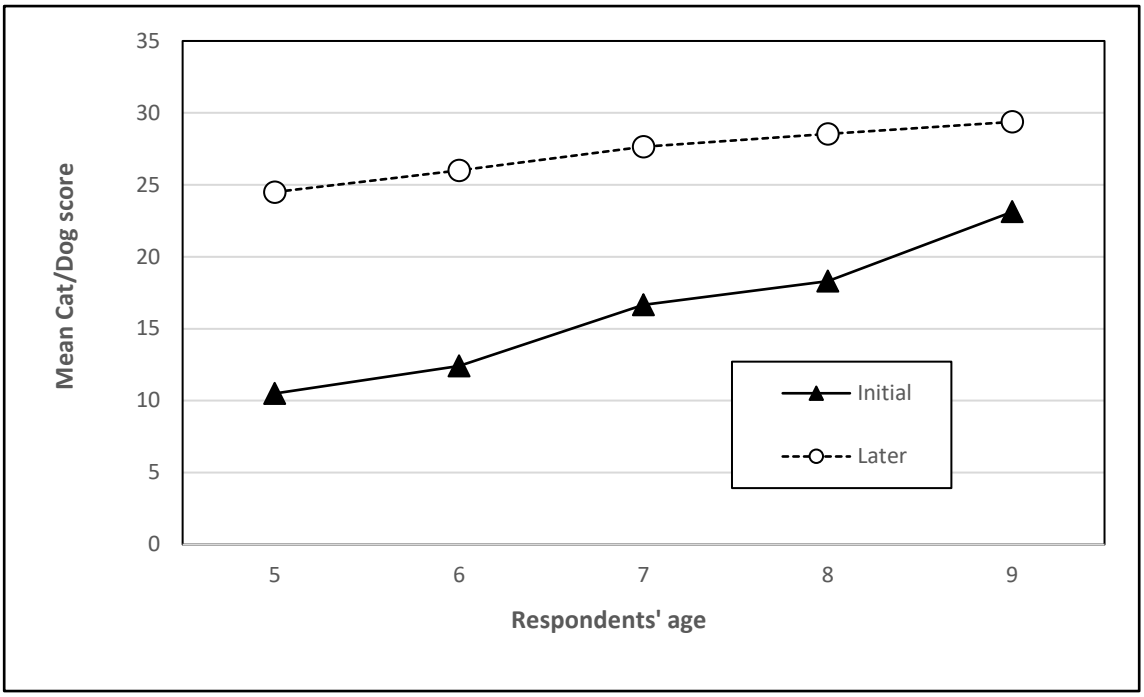
1) the within-subject repetition factor,
2) the between-subjects age factor,
3) their interaction,
4) gender
5) size.

An inspection of the means shows that the girls in the combined sample score significantly higher (mean = 21,3) than the boys (20,2). The size effect is due to the intermediate schools yielding lower mean scores (18,5) than smaller schools (22,0) as well as larger schools (21,4). This corresponds rather closely to the size effect shown for the summed scores (Cf. fig. 9).

An interesting question, then, is if this simple picture is dependent on the removal of the variance of the nonsignificant factor (Treatment). A new analysis shows that it is not; remaining results do not change if the Treat factor is removed from the analysis.

The interaction affect between Repetition and Age also needs an explanation. Firstly, figure 14 shows that both single effects are in the expected direction. The Later scores (post-intervention) are higher than the Initial ones, and both scores increase with increasing age. The increase in Initial scores, however, is stronger than the increase on Later scores, constituting the reported interaction effect.

Figure 14: Mean Cat/Dog scores in five age groups, before and after AoL interventions



These trends are quite similar to those shown in figure 6. This means that the Cat/Dog subtest is strongly influenced by respondents' age, just like the summed Yellow-Red score. More importantly, however, the AoL intervention does not significantly influence the Cat/Dog results. This also was the case with the summed scores. All in all, therefore, these ANOVA analyses show that the summed scores and the Cat/Dog subtest yield fairly similar results.

3.3.2.2 Difference scores

With a focus on the change from before to after the AoL interventions, however, also a difference score may be a convenient measure. It is obtained by subtracting the before-scores from the after-scores for each respondent. Simply put, the difference score indicates the amount of improvement from *before* to *after* the interventions for each person.

Simple difference scores have known psychometric problems, however. As Cattell (1982) pointed out, the use of difference scores has commonly been discouraged. They may serve to remove essential variance, and their reliability is less than the reliability of its two 'parent' measures. (Edwards, 2001) recognizes that "…difference scores suffer from numerous methodological problems…", and advocates polynomial regression as an alternative. Gollwitzer et al. (2014) recommend using residual change scores or latent difference score models instead; and Pedhazur and Schmelkin (1991, p. 573) prefer ANCOVA over difference scores in pretest-posttest designs.

For the purposes of the present report, however, complex psychometric arguments will be left on the side. The shortcomings of simple difference scores do exist, but to our modest explorations they will not be decisive. Rather, the easy and intuitive understanding of difference scores will be viewed as an advantage in the context of this report.

A two-sample t-test reveals that the difference between these averages is statistically significant at the 10% level (p = 0.054), indicating a slightly steeper improvement among Norwegian students. As illustrated by figure 15 at the following page, Norwegian respondents started from a slightly lower average test score during the initial testing, with both groups reaching a similar level by the later testing. Apparently, there is an important country difference after all.

But the regression table below indicates age and gender as significant explanatory variables for the difference scores. On average, younger students experienced greater improvement, and girls demonstrated a larger improvement than boys during the measured period.

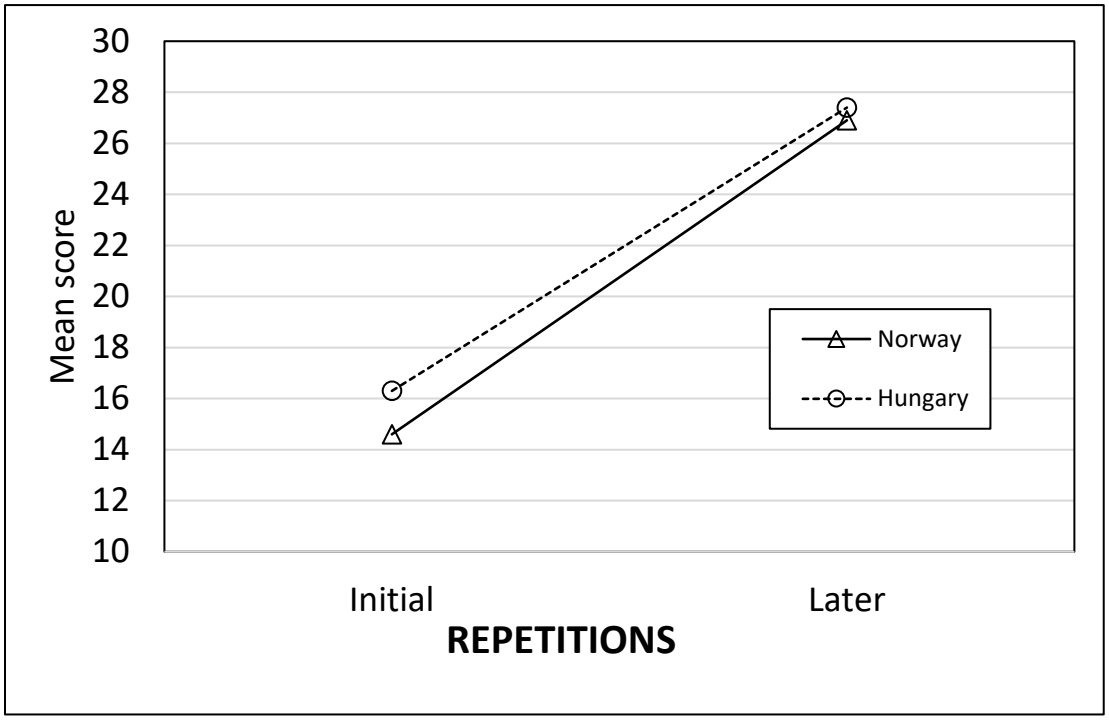
Table 4: Regression of country, treatment, age, gender, and school size variables on Cat-Dog difference scores

VARIABLES	Beta value (β)	Standard error
Country	-0.446	(0.648)
Treated	-0.558	(0.554)
Input_Age	-1.804***	(0.368)
Gender	1.433***	(0.544)
School size	0.625	(0.387)
Constant	21.24***	(2.738)
Observations	487	
R-squared	0.077	

*** p<0.01, ** p<0.05, * p<0.1

It should be borne in mind, however, that ANOVA clearly indicated that the effect of School size is not linear. As shown by figure 9, the smaller as well as the larger schools yield higher means than the intermediate. Since our regressions assume linear relationships, the non-linear relations in the data may not be detected.

Figure 15: Means of Cat-Dog scores before and after interventions in two national groups



It should also be noted, however, that the amount of variance explained (0,077 %) is rather modest. Although significant in this particular model, the confidence in age and gender as predictors of Cat/Dog score improvement should be limited.

It may seem logical that executive function improvement is negatively correlated with age. Why improvement is observed to be slightly greater among girls than among boys, however, may need an explanation.

3.3.3 The Arrows task

This task is intended to measure inhibition, much like the Cat-dog task. It may thus be reasonable to expect similar results for the two tasks.

The total sample of respondents achieved an average score improvement of 9.6 points from the initial to the later testing in the *Arrows* task. On average, Norwegian students demonstrated a development of 9.4 points, while Hungarian students showed an improvement of 10 points. A two-sample t-test reveals that the difference between these averages is not significant. ANOVA shows, however, a picture that is more complex.

3.3.3.1 Analysis of variance

A preliminary ANOVA of the Arrows data indicates that the Gender variable is the least interesting factor. Accordingly, it was dropped from further analysis.

Several variables then appear to be relevant explanatory variables for the Arrows subtest scores. The results of the central ANOVA are shown in table 5 on the following page.

Table 5: ANOVA of the Arrows subtest scores before and after interventions

Tests of Within-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Rep	2849,687	1	2849,687	232,940	0,000
Rep * Country	20,911	1	20,911	1,709	0,192
Rep * Age	331,839	4	82,960	6,781	0,000
Rep * Treated	0,428	1	0,428	0,035	0,852
Rep * Size	4,550	2	2,275	0,186	0,830
Rep * Country * Age	25,921	3	8,640	0,706	0,549
Rep * Country * Treated	34,086	1	34,086	2,786	0,096
Rep * Country * Size	52,878	2	26,439	2,161	0,116
Rep * Age * Treated	5,293	3	1,764	0,144	0,933
Rep * Age * Size	42,306	8	5,288	0,432	0,902
Rep * Treated * Size	1,370	2	0,685	0,056	0,946
Rep * Country * Input_Age * Treated	43,547	2	21,774	1,780	0,170
Rep * Country * Age * Size	3,859	2	1,930	0,158	0,854
Rep * Country * Treated * Size	0,683	1	0,683	0,056	0,813
Rep * Age * Treated * Size	65,823	5	13,165	1,076	0,373
Rep* Country * Age* Treated* Size	0,000	0			
Error(Rep)	5480,643	448	12,234		
Tests of Between-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	88276,240	1	88276,240	2950,932	0,000
Country	163,232	1	163,232	5,457	0,020
Age	401,602	4	100,400	3,356	0,010
Treated	14,343	1	14,343	0,479	0,489
Size	220,323	2	110,162	3,683	0,026
Country * Age	356,000	3	118,667	3,967	0,008
Country * Treated	10,477	1	10,477	0,350	0,554
Country * Size	114,748	2	57,374	1,918	0,148
Age * Treated	100,722	3	33,574	1,122	0,340
Age * Size	557,602	8	69,700	2,330	0,019
Treated * Size	0,991	2	0,496	0,017	0,984
Country * Age * Treated	131,904	2	65,952	2,205	0,111
Country * Age * Size	64,268	2	32,134	1,074	0,342
Country * Treated * Size	135,261	1	135,261	4,522	0,034
Age * Treated * Size	70,415	5	14,083	0,471	0,798
Country * Age * Treated * Size	0,000	0			
Error	13401,784	448	29,915		

Not only are the Rep and Age factors important again, as well as their interaction. The repetitions factor is very strong. The Age factor also is notable, as is its interaction with the Rep factor. This corresponds with findings on the Cat-Dog and summed scores already discussed. But Age also significantly interacts with Country and Size.

There also is a Rep by Country interaction, which may be more of a surprise. In addition, Country, Age and Size are strong single influences on the Arrows scores.

Again, figures may be useful to properly interpret the effects. Figure 16 shows the relationship between Repetitions and Age. Here, the Initial scores are consistently lower than the Later scores. By and large, the scores also increase with increasing age. The figure also shows, however, that the two lines are not strictly parallel. Rather, they appear to converge more as age increases. This is what the interaction effect means; the difference between Initial and Later Arrow scores diminishes with the increasing age of respondents.

Figure 16: Means of Initial and Later Arrow scores in five age groups

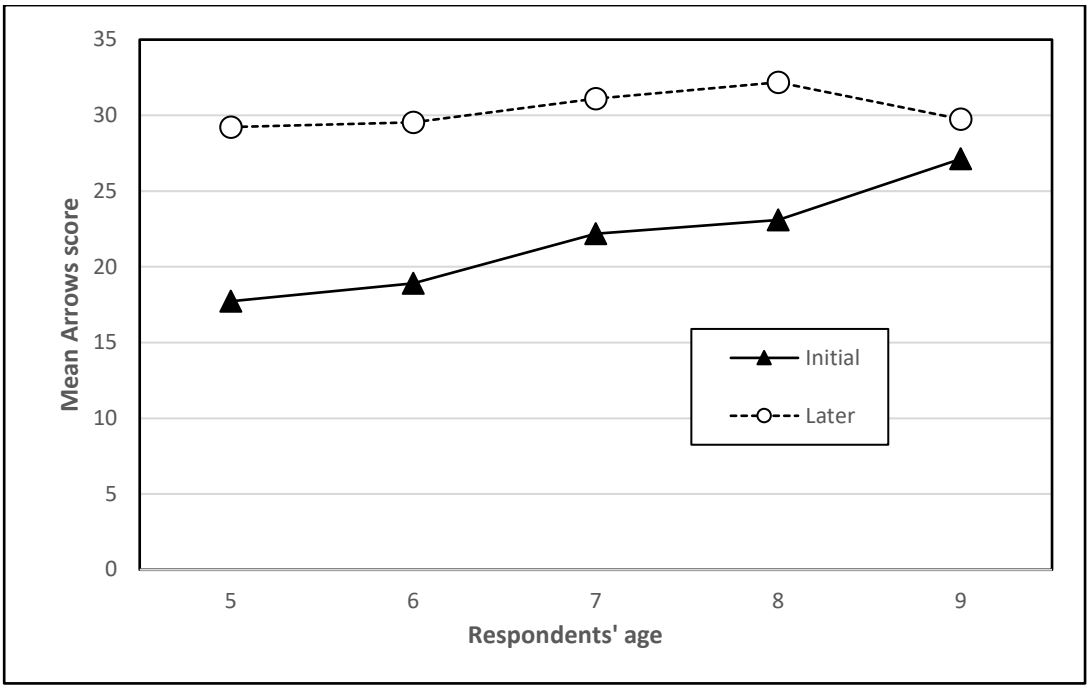
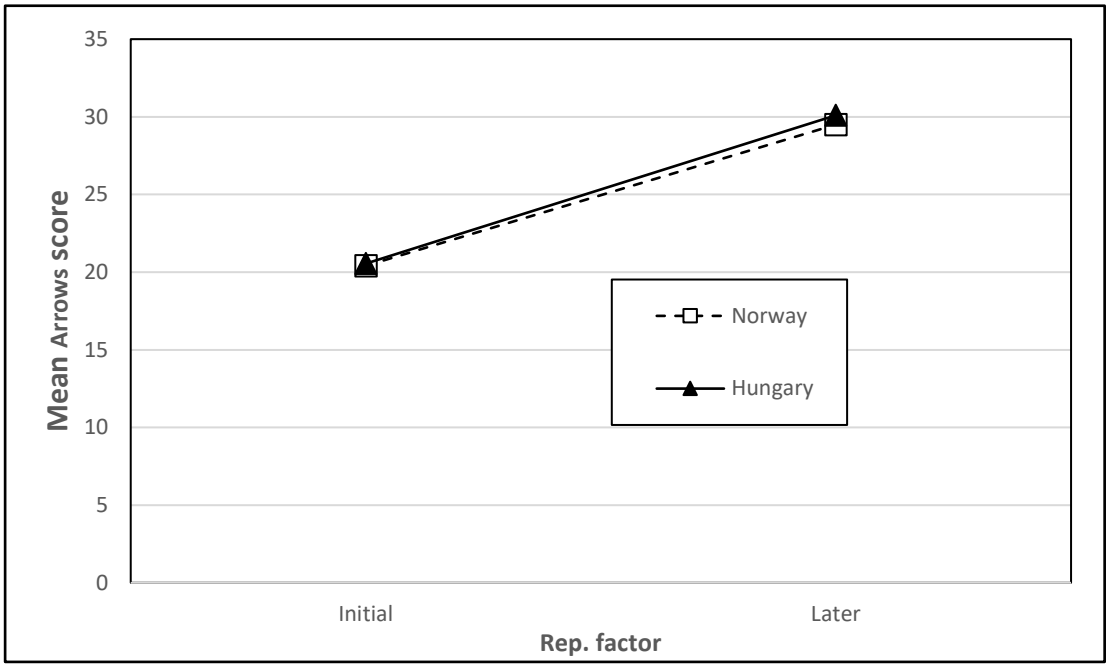


Figure 17 on the following page illustrates factors Rep and Country and the relationship between them. Clearly, the means of both countries increase from the Initial to the Later observations. This is the main effect of Rep factor. The general difference between the countries may not appear very large. It is statistically significant, however. On close inspection, the two lines are also not strictly parallel. The increase in the Hungarian mean scores is visibly larger than the increase in the Norwegian. Although not very large, also this interaction effect is significant.

Figure 17: Means of Arrows scores before and after interventions in two national groups

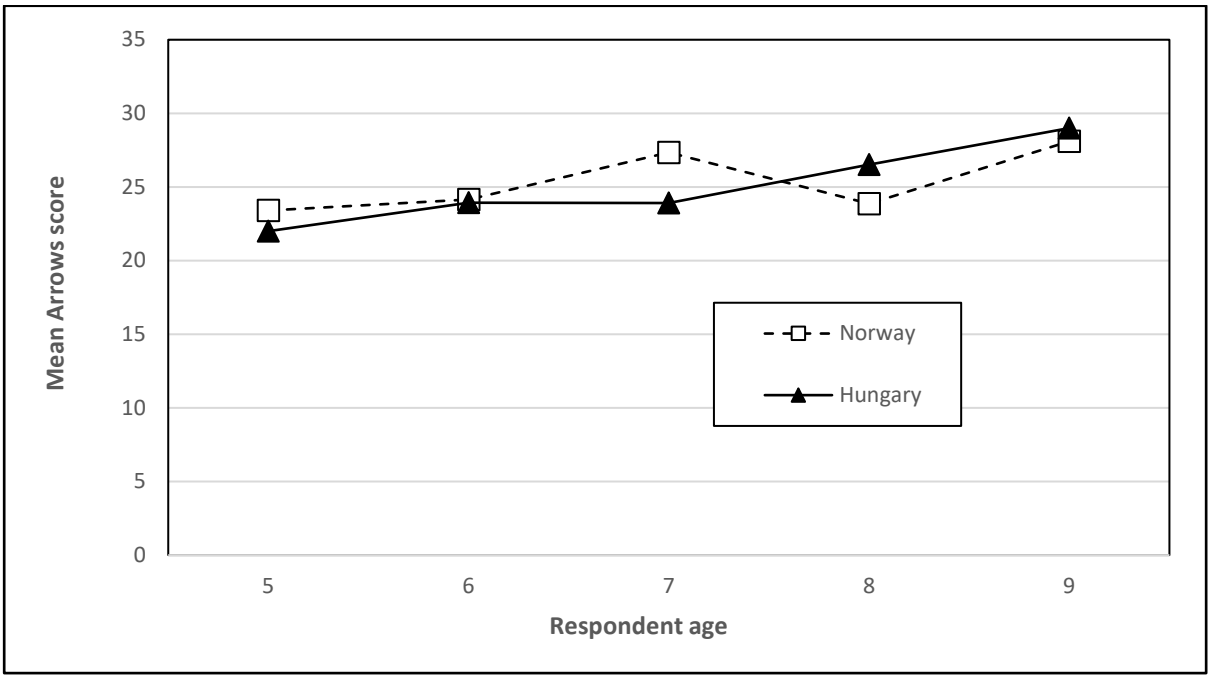


The comparison of these ANOVA results to those obtained with the summed score is relatively straightforward. The Repetition and Age factors are significant in both cases, as is their interaction. However, the Country factor (which was not significant with the summed scores and the Cat-Dog scores) arrives at 2 % with the Arrows scores.

With the Arrows scores, however, also Country interacts significantly with Age. Figure 18 on the next page shows how Arrows scores from the two countries are dependent on respondents' Age. The meaning of these variations, however, is not immediately obvious.

All in all, therefore, the ANOVA results of Arrows data prove only partly similar to those of the summed scores.

Figure 18: Means of Arrows scores at ages 5 – 9 in two national groups



3.3.3.2 Difference scores

Also here, a difference score may yield interesting insights. Applying multivariate regression analysis to the difference scores indicates that respondents' age, country and school size are significant factors, while treatment (AoL intervention) and gender are not. It should also be noted, however, that this regression model is not very strong, explaining only a limited part (0,065 %) of the Arrows variance.

Table 6: Regression of country, treatment, age, gender and school size variables on Arrows difference scores

VARIABLES	Beta value (β)	Standard error
Country	1.366**	(0.537)
Treated	-0.367	(0.459)
Input_Age	-1.626***	(0.305)
Gender	-0.117	(0.451)
School size	0.790**	(0.320)
Constant	17.20***	(2.269)
Observations	487	
R-squared	0.077	

*** p<0.01, ** p<0.05, * p<0.1

3.3.4 The Triads task

The 'Triads' task is designed to measure cognitive flexibility; the ability to change and adapt our thinking in different situations.

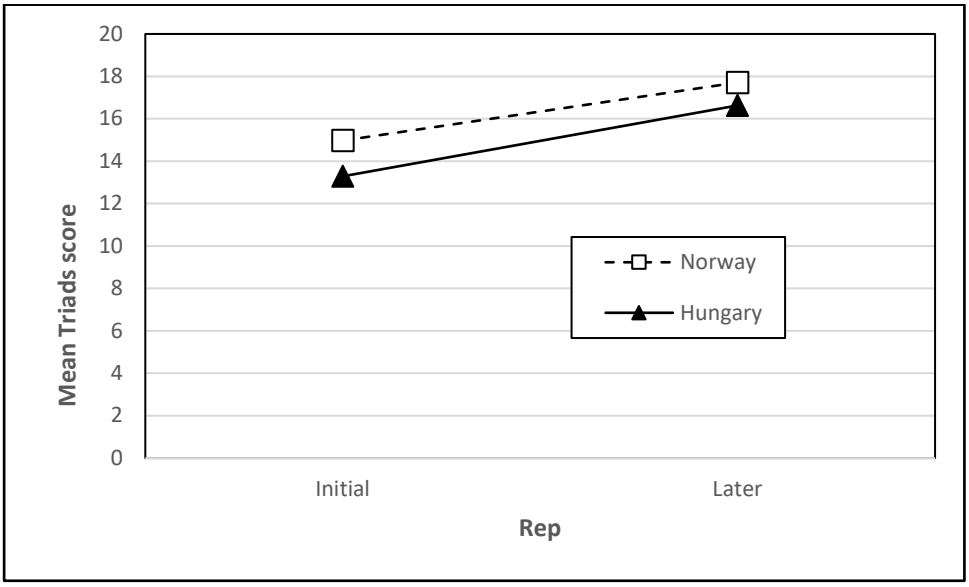
3.3.4.1 Analyses of variance

An initial ANOVA covering all variables first shows that the Treatment factor (AoL treatment vs. Control) is quite unimportant to the Triads scores. It plays no role in itself and has no significant interactions with other variables. Consequently, it will not be included in the following analyses.

Table 7 on the following page shows that the Rep factor is by far the most important predictor of the Triad means. Country and Gender are the other significant factors in this ANOVA. And for once, Rep does not interact with any other variable.

Both Country and Gender are clean main effects, exerting similar influence on both initial and later Triads measures. This may be observed in figures 18 and 19.

Figure 18: Effect of Country on the two Triads measurements



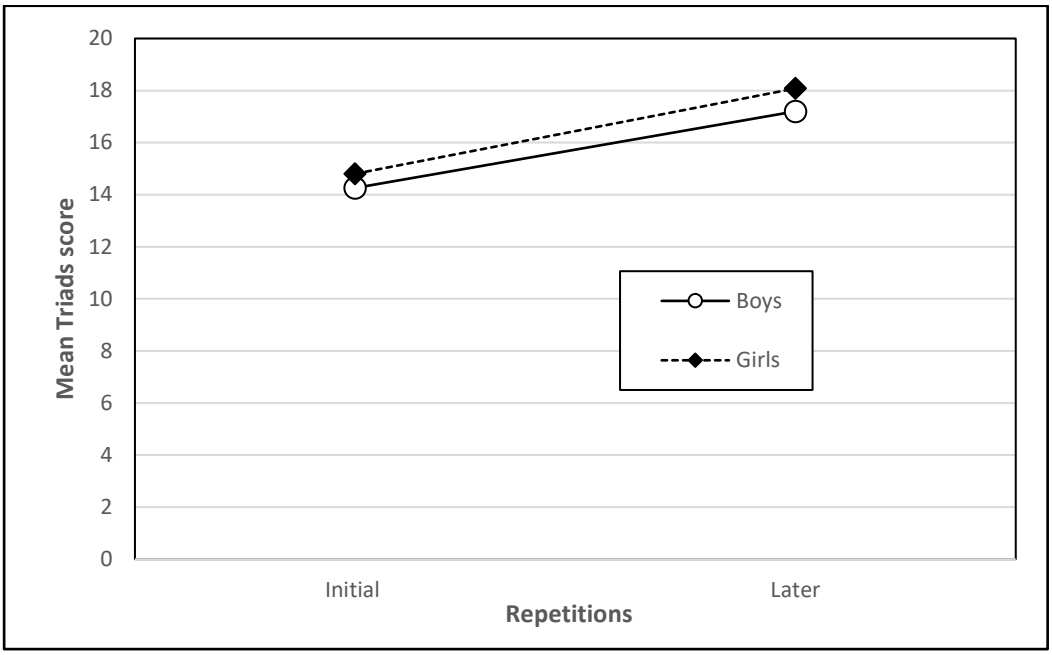
In Figure 18, the computed mean scores of the Norway group are *higher* than the means of the Hungary group. The Triads test apparently is easier for Norwegian children than for the Hungarian.

Please note, however, that these means are based on a modified population marginal mean. They are thus computed after correcting for the effects of all variables in the analysis of variance. Simpler group means, computed independently from the raw data of each group, do not necessarily show the same picture.

Table 7: ANOVA of the Triads subtest scores before and after interventions

Tests of Within-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Rep	287,628	1	287,628	27,932	0,000
Rep * Country	6,856	1	6,856	0,666	0,415
Rep * Age	13,851	4	3,463	0,336	0,854
Rep * Gender	1,706	1	1,706	0,166	0,684
Rep * Size	27,908	2	13,954	1,355	0,259
Rep * Country * Age	49,386	4	12,346	1,199	0,311
Rep * Country * Gender	6,149	1	6,149	0,597	0,440
Rep * Country * Size	7,034	2	3,517	0,342	0,711
Rep * Age * Gender	34,062	4	8,515	0,827	0,508
Rep * Age * Size	52,869	8	6,609	0,642	0,743
Rep * Gender * Size	16,666	2	8,333	0,809	0,446
Rep * Country * Age * Gender	32,884	2	16,442	1,597	0,204
Rep * Country * Age * Size	0,450	2	0,225	0,022	0,978
Rep * Country * Gender * Size	8,062	2	4,031	0,391	0,676
Rep * Age * Gender * Size	12,895	5	2,579	0,250	0,940
Rep* Country * Age* Gender* Size	2,509	1	2,509	0,244	0,622
Error(Rep)	4561,791	443	10,297		
Tests of Between-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	36079,286	1	36079,286	3102,744	0,000
Country	99,646	1	99,646	8,569	0,004
Age	41,615	4	10,404	0,895	0,467
Gender	79,522	1	79,522	6,839	0,009
Size	38,755	2	19,378	1,666	0,190
Country * Age	28,095	4	7,024	0,604	0,660
Country * Gender	3,998	1	3,998	0,344	0,558
Country * Size	17,630	2	8,815	0,758	0,469
Age * Gender	82,792	4	20,698	1,780	0,132
Age * Size	107,072	8	13,384	1,151	0,328
Gender * Size	58,925	2	29,462	2,534	0,081
Country * Age * Gender	97,056	2	48,528	4,173	0,016
Country * Age * Size	27,714	2	13,857	1,192	0,305
Country * Gender * Size	112,192	2	56,096	4,824	0,008
Age * Gender * Size	57,930	5	11,586	0,996	0,419
Country * Age * Gender * Size	21,576	1	21,576	1,856	0,174
Error	5151,287	443	11,628		

Figure 19: Effect of Gender on the two Triads measurements



In Figure 19, girls score higher than boys on both initial and later measurements. This suggests that the Triads subtest *generally* is easier for girls than for boys.

If this is correct, it may well deserve some attention, since gender differences were not found with the summed Yellow-Red scores. May this be interpreted as a potentially interesting finding? Could it, e.g., be an argument against the present practice of simply summing the results of all four subtests into one common score?

The Rep effect of the Triads scores matches well with what was consistently found with the summed scores; later measures are commonly higher than the initial.

But the Triads scores prove different from the summed scores in several respects. Firstly, the Gender difference only holds only for the Triads scores, not for the summed scores. More importantly, however, respondents' Age – which was a significant factor with the summed scores – is not important with the Triads data. These differences are not trivial and may suggest that the Triads subtest and the summed Yellow/Red score are measuring different things.

3.3.4.2 Difference scores

In the regression table (Table 8), Country is the only significant predictors of the Triads difference scores. This was also the case in the ANOVA in table 7, providing some agreement. The ANOVA, however, also pointed to Gender as an important predictor. This is not significant with the difference scores.

The regression model's explanatory power is also very limited (0.025). All in all, therefore, the regression on the Triads output scores has little to offer beyond the ANOVA results.

Table 8: Regression of country, treatment, age, gender and school-size variables on Triads difference test scores

VARIABLES	Beta value (β)	Standard error
Country	1.078**	(0.486)
Treated	0.678	(0.415)
Input_Age	0.211	(0.276)
Gender	0.286	(0.408)
School size	0.351	(0.290)
Constant	-1.331	(2.053)
Observations	485	
R-squared	0.025	

*** p<0.01, ** p<0.05, * p<0.1

3.3.5 The Bindings task

The binding test is designed to capture the function of working memory: the ability to hold and manipulate information in the mind.

3.3.5.1 Analysis of variance

A preliminary ANOVA covering all relevant variables suggests that the Treatment factor plays no role to the Bindings scores. With the Treatment factor therefore not included, an ANOVA of the Bindings data produces the results shown in table 9 on the following page.

Factors Rep and Age will be examined first, since both appear as significant simple factors. As shown in figure 20, the mean scores generally increase with higher age (Age factor), and Initial scores are generally lower than the Later ones (Rep factor).

Figure 20: The interaction effect of factors Rep and Age

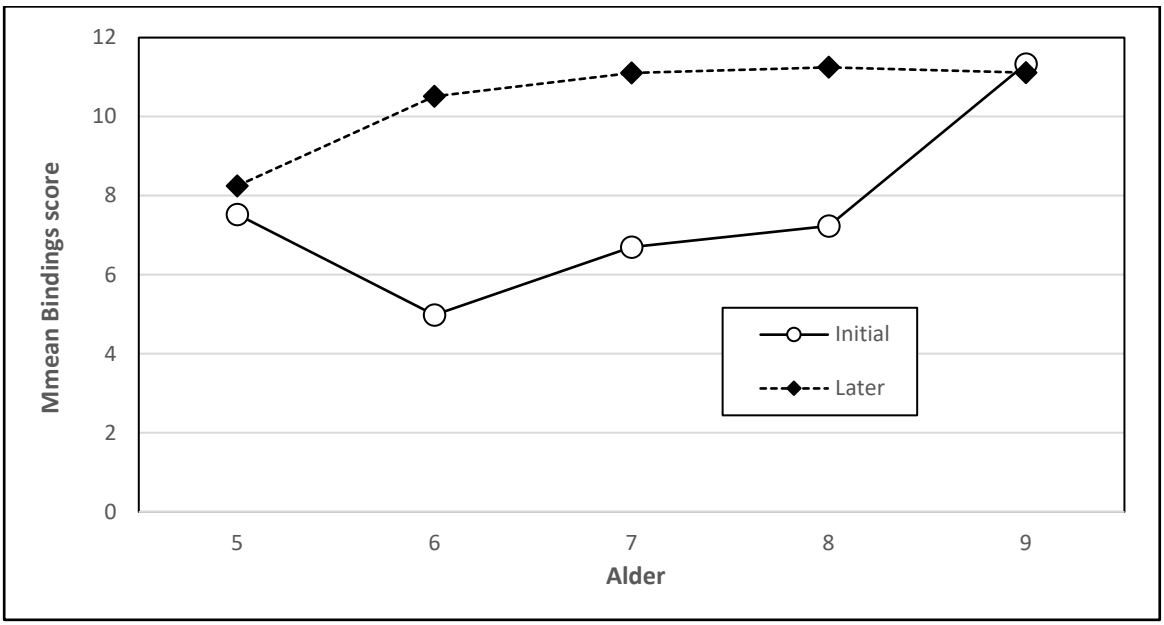


Table 9: ANOVA of the Binding subtest scores before and after interventions

Tests of Within-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Rep	358,669	1	358,669	56,491	0,000
Rep * Country	4,866	1	4,866	0,766	0,382
Rep * Age	155,447	4	38,862	6,121	0,000
Rep * Gender	0,530	1	0,530	0,084	0,773
Rep * Size	15,921	2	7,961	1,254	0,286
Rep * Country * Age	78,566	4	19,641	3,094	0,016
Rep * Country * Gender	2,662	1	2,662	0,419	0,518
Rep * Country * Size	1,867	2	0,934	0,147	0,863
Rep * Age * Gender	18,619	4	4,655	0,733	0,570
Rep * Age * Size	25,550	8	3,194	0,503	0,854
Rep * Gender * Size	8,121	2	4,061	0,640	0,528
Rep * Country * Age * Gender	0,115	2	0,058	0,009	0,991
Rep * Country * Age * Size	8,240	2	4,120	0,649	0,523
Rep * Country * Gender * Size	7,799	2	3,900	0,614	0,542
Rep * Age * Gender * Size	31,548	5	6,310	0,994	0,421
Rep * Country * Age * Gender * Size	0,736	1	0,736	0,116	0,734
Error(Rep)	2825,375	445	6,349		
Tests of Between-Subjects Effects					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	11436,776	1	11436,776	1082,609	0,000
Country	4,135	1	4,135	0,391	0,532
Age	254,762	4	63,690	6,029	0,000
Gender	39,753	1	39,753	3,763	0,053
Size	110,580	2	55,290	5,234	0,006
Country * Age	22,524	4	5,631	0,533	0,712
Country * Gender	0,001	1	0,001	0,000	0,992
Country * Size	66,347	2	33,174	3,140	0,044
Age * Gender	16,246	4	4,062	0,384	0,820
Age * Size	193,542	8	24,193	2,290	0,021
Gender * Size	15,618	2	7,809	0,739	0,478
Country * Age * Gender	52,649	2	26,324	2,492	0,084
Country * Age * Size	31,947	2	15,973	1,512	0,222
Country * Gender * Size	47,291	2	23,646	2,238	0,108
Age * Gender * Size	104,113	5	20,823	1,971	0,082
Country * Age * Gender * Size	0,076	1	0,076	0,007	0,933
Error	4701,019	445	10,564		

The significant interaction effect between factors Rep and Age is more complicated, however. Clearly, the two plotting lines (one for the Initial and one for Later observations) follow rather different patterns with increasing age. The line for Initial measures starts with a rather high mean at five years. It then *drops* at six years but proceeds with a regular rise from six to nine years. The line for the Later

measures, however, increases already from its beginning at five years and appears to approach a limiting 'ceiling' effect at seven or eight years.

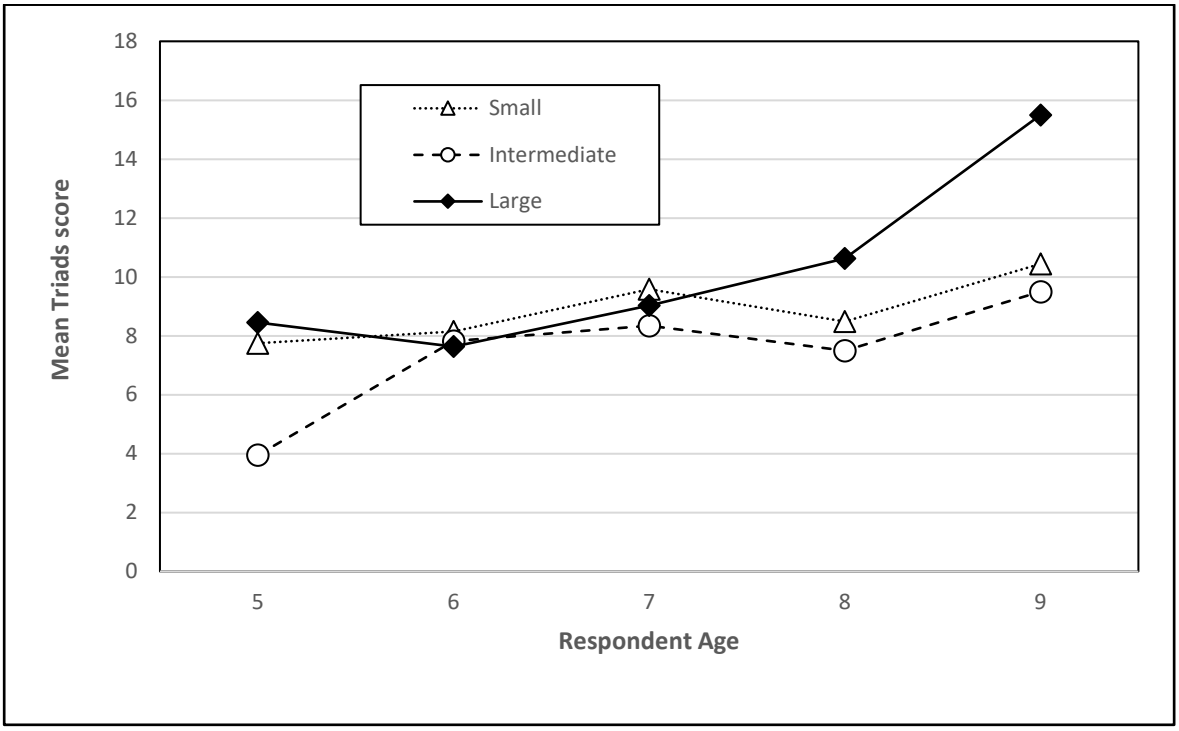
The two different trajectories show what the interaction effect is all about: The effect of Age on Initial scores is very different from its effect on Later scores. Unfortunately, the meaning of this complex relationship is less clear. It should be borne in mind, however, that the number of 5-year- and 9-year-olds in the sample is very limited (cf. fig 2). This leaves both groups exposed to random fluctuations and less trustworthy data.

The Size variable also plays a role here. It not only constitutes a significant main effect but also interacts with Age in a potentially interesting way. The meaning of this may be seen in figure 21.

The first main effect (of Age) may not be easily observable. By and large, however, the scores get higher with increasing age. The second main effect (of Size) is that means from the intermediate schools are generally lower than the means from smaller and larger schools.

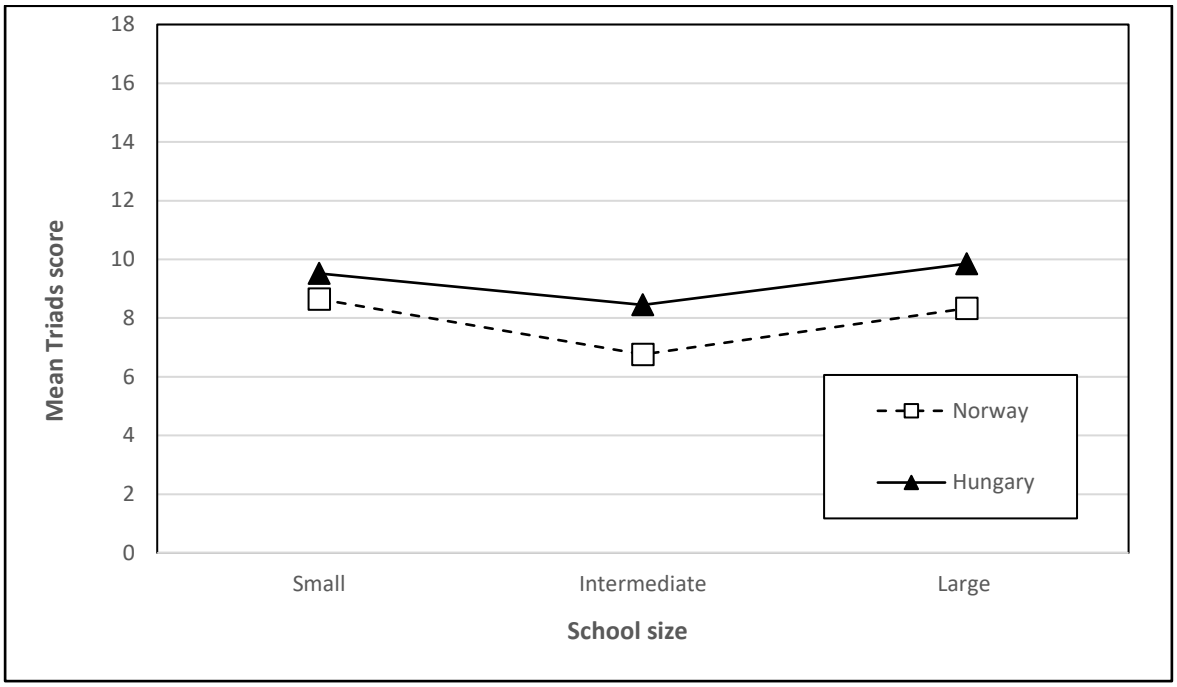
One way of describing the interaction effect would be to say that the relationship between the three school sizes depends on the age of the respondents. Or perhaps more precisely: Intermediate schools yield unexpectedly low scores at age 5, while large schools get their best scores at ages 8 and 9. Hopefully, experienced teachers will be able to make sense of these complex relationships.

Figure 21: The effect of School size, Age, and their interaction on Binding scores



The last influence to consider is the interaction between factors Country and Size, shown in figure 22 on the following page. The main effect of size is clear; intermediate schools score less than others. The interaction effect is that the difference between Countries is smaller in small schools than in the intermediate and large.

Figure 22: The effect of School size, Country and their interaction on Binding scores



It may be instructive to compare the results from the Binding subtest to the 'standard' summed scores of the Yellow-Red test. First, neither Binding nor Summed scores are affected by the Treatment (AoL Intervention vs. Control) factor. Second, the Country factor is not important to the Summed scores. On the Binding scores, however, the Hungarian children score significantly better than the Norwegian. Third, while both Initial and Later scores improve with increasing age with the Summed scores; Binding scores show a more complicated relationship to Age.

3.3.5.2 Difference scores

All factors may of course be considered in isolation, i.e. without taking into account the variance due to other factors. If this is done with the country factor, the total sample of respondents achieved an average score improvement of 4.6 points in the Binding task from the initial to the later testing. Norwegian students showed an average gain of 4.4 points, while Hungarian students demonstrated an improvement of 5 points. A two-sample t-test of difference scores in the two countries, however, reveals no statistically significant difference. That suggests that there is no difference in the magnitude of development.

Still, the analysis of the difference scores from the Binding task (cf. table 10) yield results that are similar to those of the Arrows task. The change in test scores is not only negatively correlated with age. Country is also a significant factor, and Hungarian respondents show greater improvement than the Norwegian.

This significant country difference corresponds to an interaction effect in the ANOVA, but the ANOVA only indicated a main effect of Country. The general Gender difference in the ANOVA is also not replicated in the difference scores. The regression equation only explains a very small part of the variance (2 %), however; and may thus not deserve further attention.

Table 10: Regression of country, treatment, age, gender and school size variables on Bindings difference scores

VARIABLES	Beta value (β)	Standard error
Country	0.912**	(0.402)
Treated	-0.0297	(0.343)
Input_Age	-0.590***	(0.228)
Gender	0.0259	(0.337)
School size	0.135	(0.240)
Constant	7.001***	(1.698)
Observations	487	
R-squared	0.021	

3.4 Summary of key findings

In the table below, significant effects of the five ANOVAs are listed. To preserve readability, second-order interaction effects are not included.

Table 11: ANOVA of summed scores and subtests; significant main and first-order interaction effects

ANOVA	Rep	Cty	Age	Trm	Gdr	Size	RxC	RxA	RxT	RxS	CxA	CxT	CxS	AxT	AxS	TxS
Sumscore	√		√			√		√			√	√	√		√	
Cat-Dog	√		√		√	√		√								
Arrows	√	√	√			√		√			√				√	
Triads	√	√			√											
Binding	√		√			√		√					√		√	

Judging from the joint pattern in the table, three tentative conclusions are tempting. Firstly, the Treatment factor is clearly not an important one. This main effect is not significant in any of the five analyses. And the interactions Rep x Treatment, Age x Treatment or Treatment x Size also are not. The most important one, of course, is the R x T interaction, testing the hypothesis that the AoL treatment is different from the Control condition. Notably, none of the five (RxT) analyses indicate any effect of the AoL intervention.

Secondly, the Rep factor is universally important; all measures – including the summed scores – increase from Initial to Later observations. However, a closer examination of the subtests indicates that the Arrows and Cat-Dog tasks showed stronger improvement over time compared to the Binding and Triads tasks. This suggests that different aspects of executive functions may have responded differently to the intervention.

Thirdly, factors Age and Size as well as the Rep by Age interaction are usually significant. The Triads subtest is an exception, however. It sets itself apart by not conforming to this general pattern of relations and is only influenced by the Repetitions and Country factors.

Analyses of variance of the findings from the Yellow-Red test thus suggest that while executive function (EF) scores were improved over time, this change was not due to Art of Learning (AoL) interventions (RxT). Rather, age-related development (RxA) appears to exert a general influence.

In table 12, the results of regression analyses of the difference scores of the four subtests are shown. Since difference scores cover the difference between the two Rep conditions (Later observations – Initial observations), the Rep factor is not available for these analyses. The factors Country, Age, Treatment, Gender, and (school) Size were thus entered into the regression equation, using the Enter method. In these analyses, some interesting results appeared.

Table 12: Linear regressions of difference scores of four subtests

Diff. scores	Cty	Age	Trm	Gdr	Size
Cat-Dog	√	√			√
Arrows	√	√			
Triads	√				
Binding	√	√			

The most important finding in table 12 is that the Treatment effect does not prove to be a significant predictor for the difference scores of any of the four subtests. Exposure to the Art-of-learning intervention does not influence the difference scores. This confirms the central ANOVA finding in table 11. Not only was the Replication by Treatment interaction (RxT) insignificant for the summed scores – it also was for all the four subtests.

Among the five predictor variables, Country proved significant to all subtests of table 12. In table 11, however, only the summed score had a significant Country by Treatment (CxT) effect. It may nonetheless be noted that a main effect of Country was found for the Arrows and Triads subtests.

Moreover, Age was significant to all subtests except Triads. This matches rather well with the findings of table 11. Also in accordance with that table, the Treatment factor was not significantly related to any of the subtests.

But all in all, the regressions on difference scores mainly confirm what had already been shown by the ANOVAs. The change of scores from Initial to Later observation in the ANOVAs – as well as the difference scores in the regressions – are not influenced by the Treatment factor. Rather, Age appears to be a more potent influence. Being closely related to the very purpose of the project as well as its design, this part of the results clearly deserves attention.

However, a number of other interesting relationships may be identified in the data. Country, e.g., appears as a powerful predictor in several instances. So does School size and Gender, and several interaction effects between the independent factors. While not directly relevant to the project's central hypotheses, this may yield useful knowledge for closer evaluations of the project and its procedures, and for further work in this field.

4. Discussion

The findings of this study suggest that executive function improvements, as measured by the Yellow-Red test, occur naturally with age rather than as a direct consequence of art-based interventions. While previous literature has emphasized the potential of creative pedagogies to enhance cognitive flexibility and inhibitory control (Diamond, 2013; Németh, 2023), the present study finds no statistically significant effect of the Art of Learning (AoL) program on overall executive function scores.

The first part of this chapter will thus discuss the relationships between the Yellow-Red test scores, age, and executive functions. In the second part, the information gained from numerous supplementary analyses will be evaluated. Understanding the influence of a wider range of potentially influential variables may be useful to future adjustments and development in the project.

4.1 Executive functions, Yellow-Red test scores and age

The first hypothesis of the project implies that the executive functions (as measured by Yellow-Red test scores) in the Treatment group will improve more than in the Control group. The findings of the present study do not confirm this. Executive function improvements is about the same in the group subjected to art-based interventions and the control group where these interventions are not used. Data also show age to be an important influence on the Yellow-Red scores, confirming the project's second hypothesis.

It may be worth noting that also a Norwegian pilot study (Kleiven et al., 2022) reached the same conclusion. Other data from the same pilot study do support the hypothesis, however. Using teacher interview data and the global executive composite (GEC) score and the behavioral regulation index (BRI) of the BRIEF (Gioia et al., 2000) inventory; Andersen et al. (2019) shows that the progress of the AoL group is stronger than that of the control group. Their conclusion, therefore, is that "the AoL program shows promising effects on behavioral self-regulation (BRI) improvement in children aged 6– 9 years as reported both from teacher rating scales and interviews". The Andersen et al. (2019) paper makes no mention of their Yellow-Red test results of their study, however.

The data basis of the present study, of course, is new and different. But again, the improvement from initial scores (before the intervention) to later scores (after the intervention) is about the same in the AoL and the control groups. This holds not only for the summed score of the test, but also for three out of its four subtests. Our Yellow-Red data thus does not support the belief that the AoL experiences in the project improve children's executive functions.

4.1.1 The test and executive functions

The unexpected lack of an AoL treatment effect raises important questions about the measurement of EF in arts-based education and highlights the need for further research into alternative assessment methods and the broader impacts of creative pedagogy.

Put more simply, the negative result may lead to questions about the test and its fundamental concepts. First of all, is this test trustworthy in this context? Recent publications on the Yellow-Red (Rosas-Días et al., 2019) (Rosas et al., 2022) offer rather convincing accounts of the properties and

qualities of the test. Its close relationship to age also lends some credibility. This leaves little room for distrusting it.

Second, does the Yellow-Red test really measure what the project needs to know? When Diamond (2013) defined executive functions, they were clearly viewed as a consisting of partly different processes. Inhibition, working memory, and cognitive flexibility, e.g., are all 'executive functions'. They are also different and distinguishable phenomena, however, and the four subtests of the Yellow-Red are indeed intended to cover partly different functions.

Yet, our negative finding may be at some variance with previous literature that has emphasized the potential of creative pedagogies to enhance cognitive flexibility and inhibitory control (Diamond, 2013; Németh, 2023). Further efforts at developing and refining the central concepts of the project may thus be interesting next steps. With some luck, this could also imply alternate tests or new ways of measuring relevant executive functions.

4.1.2 The Age and Country variables: useful proxies?

Statistically, the importance of Age to the Yellow-Red scores is rather clear. The causal meaning of this variable, however, may be less obvious. Age in itself do not influence the test scores, and may thus not be a causal variable in our context.

A number of other variables mirrors the age variable, however, being highly correlated with it. Maturation and general learning, e.g., naturally coincides with increasing age. When children grow, both physical development and learning experiences take place in their lives, gradually improving executive functions and problem-solving capacity year by year. These processes may thus be actual influences on the executive processes and also contribute to improved models of cognition. Still, they should not be confused with the Age variable, which simply measures the passage of time.

When the present project controls for children's age, then, it is because it coincides with other variables that are expected to influence executive functions. Age may be a proxy for a larger array of variables.

Maturation has already been mentioned as a potentially disturbing factor. It is, of course, closely related to children's age. Through physical growth, children normally develop increasing capacities, including their problem-solving ability. In principle, some improvement in performance should thus be expected on all retests. It should be borne in mind, however, that maturation may be a causal variable, by providing a direct, partial explanation of children's increasing competence. Age may be an adequate statistical proxy, but its relationship to executive functions is less direct. Additional concepts are needed to form a causal model of the processes involved.

Another interesting concept is general or indirect learning. Throughout life, children gradually learn general strategies of handling problems and become better at it. And again, skills and dexterities acquired elsewhere may improve the specific Yellow-Red performance. Becoming more familiar with testing, video games, screens and keyboards may be an example of useful general learning.

But specific and relevant learning is even more likely to be important here. In their initial Yellow-Red exposures, children do learn, e.g., what happens in the test, how to meet its different challenges, and which mistakes should be avoided. Naturally, this may yield a better performance in their final testing, where much is known in advance.

The Hawthorne effect may be another example of a confounding variable. When respondents are aware of being watched, they commonly react by improving their performance[7]. For the Yellow-Red,

[7] This effect is *not* a universal truth. It may occur, however, given the right conditions.

this could mean children that are highly motivated in general; thus performing very well in both pre- and postintervention tests.

Maturation, however, probably works together with general and specific learning to yield a better performance in the second Yellow-Red test than in the initial one. In the face of this combined effect, a less powerful AoL effect may be masked or hidden.

A better approach to the Age variable, may be not to view it as a causal variable. Age itself does not influence other the Yellow-Red scores; it impacts executive functions through its close correspondence with maturation and learning. The coincidence of AoL intervention and improved scores should not be confused with causality. The statistical correlation needs a different account of its causality.

A similar judgment may be applied to the variable of School size. Instead of assuming that School size itself can explain different test results, the focus should be on finding other variables. What are the problems with intermediate schools? Are there, e.g., worse social relations in their classes, less motivated teachers, or less engaged parents? And what are the advantages of larger and smaller schools; do they have better buildings, more resources, or better educated parents?

School size is also not a causal variable; its effect comes from its correlations with other phenomena that matter. Small, intermediate and large schools may, e.g., have different access to resources and different proportions of disadvantaged children. They may also not recruit students or teachers equally well or have different practical options for children in their neighborhood. Quite likely, the intermediate schools in the project have some relative disadvantages that limit student motivation and performance.

In this context, the interaction between Size and Country may be interesting. Could it be a signal that causal factors play different roles in the two countries? Or more generally; why do Norwegian children in small schools score better than others in Norway, while Hungarian children score best in intermediate and large schools? Are, e.g., underprivileged children common in the small schools of Hungary, but rather rare in comparable Norwegian schools? Explanations should be tied to causally relevant concepts, not to general statistical variables. Correlations are not to be confused with causality.

4.2 Supplementary analyses

The analysis of individual subtests reveals more complex patterns, some of which may deserve further attention. The four subtests are not only intended to measure different facets of the executive functions. They also contribute differently to the common summed score (Cf. section 3.3.1). More importantly, scores of subtests Cat-Dog and Arrows are both intended to tap inhibition. These scores increase more from Initial to Later than do the scores of Triads and Binding. In future discussions of alternative concepts or tests for AoL projects, therefore, the process of inhibition should be kept in mind.

It may also be of interest that the four subtests commonly relate to background variables in different ways. Gender, e.g., is a significant predictor only to the scores of Cat-Dog and Triads. On both subtests, girls score significantly higher than boys[8]. And the Country variable makes a difference only to the Arrows and Triads subscales. In both cases, the Hungarian mean scores are higher than the Norwegian.

[8] The Gender difference is also very close to significance on the Binding scale (p = 0,053).

These findings may indicate that executive function development is not quite the same in boys and girls, or with Norwegian and Hungarian children. Rather, the results of subtests may be different, and shaped by broader cultural and educational influences. This perhaps suggests that future studies should adopt a more granular approach to intervention efficacy, moving beyond aggregated scores to examine how different cognitive domains respond to arts-based learning.

Clearly, the Treatment factor is not significant; most score means improve in the Control group as well as in the "Experimental" group. But results also show that the Repeat procedure is more (or less) effective under certain circumstances. The repetitions (Initial –> Later) may, e.g., be more important to only parts of the sample (gender, nationality, age, school size). They may also be more relevant to or to certain subtests than to others.

The results thus may suggest that more tailored approaches be considered for future interventions. Specifically, focus could then be on, e.g. :

- Younger students, where intervention effects might be more pronounced.
- School environments, ensuring resources and engagement are sufficient in different school sizes.
- Cultural contexts, as differences between Norway and Hungary suggest broader educational and cultural influences.

More thought could also be given to the problem of influences that run parallel to age effects. As shown in section 4.1.2, other important things happen with the progress of time or age. Children grow and mature naturally and simultaneously acquire new knowledge and skills – all of which may well be influences on their executive processes that are stronger than the project's Art of Learning interventions.

To avoid confounding age-related with AoL-related effects, then, searching for alternative concepts and measures of effect may be advisable. An interesting focus could be on person characteristics that are relatively stable, not so easily subject to gradual development or change. Personal values or central attitudes towards arts, culture or education may perhaps serve as examples of this. Hopefully, selecting attitudes that also have known and tested measurement properties would increase the chances of getting hard data to demonstrate positive effects of AoL interventions.

References

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

Cattell, R. B. (1982). The clinical use of difference scores: Some psychometric problems. . *Multivariate Experimental Clinical Research*, *6*, 87-98.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*, 135-168. https://doi.org/10.1146/annurev-psych-113011-143750

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, *4*(3), 265-287.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *Behavior Rating Inventory of Executive Function – Professional manual*.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2015). *Behavior Rating Inventory of Executive Function®, Second Edition (BRIEF®2)*. PAR Inc.

Gollwitzer, M., Christ, O., & Lemmer, G. (2014). Individual differeneces make a difference: On the use and the psychometric properties of difference scores in social psychology. *European Journal of Social Psychology*, *44*, 673-682.

Hendrickson, N. K., & McCrimmon, A. W. (2019). Test review of BRIEF2. *Canadian Journal of School Psychology*, *34*(1), 73-78.

Hundevadt, M. O. (2022). *"Kunsten å lære. Ei diskursanalyse av tankemønstera som ligg til grunn for Kunsten å lære-prosjektet 2021-2024".* [Master's thesis, University of South-Eastern Norway]. Kongsberg.

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"*. O. fylkeskommune.

Kaderják, A. (2024). Final Report on the Y/R test results of the Art of Learning Program in Hungary. In I. T-Tudok (Ed.). T-Tudok, Inc., Budapest.

Kleiven, J., Andersen, P. N., & Håkansson, U. (2022). *The "Yellow-Red test": A closer look at data from a longitudinal pilot project with Norwegian children* (Skriftserien, Issue. H. i. Innlandet.

Németh, S. (2023). The Art of Learning: Developing Executive Functions of the Brain by Creative Pedagogy. In Z. Molnár-Kovács, H. Andl, & J. Steklács (Eds.), *Új kutatások a neveléstudományokban 2022* (pp. 220-229). MTA Pedagógiai Tudományos Bizottság.

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

Rosas-Días, R., Espinoza, V., & Garolera, M. (2019). *Intercultural evidence of a Tablet based executive functions test for children between 7 to 10 years* EARLI (European Association for Research on Learning and Instruction) conference, Aachen.

Rosas, R., Espinoza, V., Martinez, C., & Santa-Cruz, C. (2022). Playful Testing of Executive Functions with Yellow-Red: Tablet-Based Battery for Children between 6 and 11 *Journal of Intelligence*, *10*, 25. https://doi.org/https://doi.org/10.3390/jintelligence10040125

Santa-Cruz, C., & Rosas, R. (2017). Cartografía de las Funciones Ejecutivas. *Studies in Psychology/Estudias de Psicología*, *38*(2), 284-310.

Zágon, M., & Németh, S. (2022). Experiences of attending lessons of the Creative Partnerships and the Art of Learning programmes. Monitoring report. In I. T-Tudok (Ed.). T-Tudok, Inc., Budapest.

Østern, T. P., Dahl, T., Strømme, A., Petersen, J. A., Anna-Lena, Ø., & Selander, S. (2019). *Dybde//læring – en flerfaglig, relasjonell og skapende tilnærming*. Universitetsforlaget.

The *Art of Learning* project in Norway and Hungary wished to see if art experiences in school improve children's executive functions. For this purpose, the Yellow-Red test was conducted at the start and the end of the project. After correction for age differences, however, children participating in the AoL project and a control group got the same results. This probably means that

1. The test does not measure executive functions in a satisfactory manner or

2. Participation in the project does not influence the children's executive functions.

The report also summarizes parts of the data in more detail, since it may also be relevant to further discussions on the practical procedures of the project and its future.

Art of Learning-prosjektet i Norge og Ungarn ønsket å se om kunstopplevelser på skolen forbedrer barnas eksekutive funksjoner. Derfor ble Yellow-Red testen administrert ved begynnelsen og slutten av prosjektet.

Etter korreksjon for aldersforskjeller fikk imidlertid barna som tok del i prosjektet og en kontrollgruppe like resultater. Dette betyr trolig at

1. Testen måler ikke eksekutive funksjoner på en tilfredsstillende måte eller

2. Deltakelse i prosjektet påvirker ikke barnas eksekutive funksjoner.

Rapporten oppsummerer også deler av data mer detaljert, siden de også kan være relevante for videre diskusjoner om de praktiske prosedyrene i prosjektet og dets framtid.

University of Inland Norway

inn.no