Adult Online Assessments of Arithmetic, Vocabulary, Reasoning, and Math Self-Assessment

An Evaluation of the Feasibility, Reliability and Validity.

Madelief Kuijper[1], Britt Min[2], Alexandra Starr[1], Bruno Sauce[1] & Elsje van Bergen[1,2] (2025)

Affiliations:

[1] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2] University of Oslo, Oslo, Norway

Corresponding author:

Madelief Kuijper

m.i.kuijper@vu.nl

Table of Contents

Table of Contents ... 2
Foreword ... 4
Acknowledgements ... 4
Abstract ... 5
Pilot I ... 6
Introduction ... 6
Methods ... 6
Participants ... 6
Procedure ... 6
Test descriptions ... 7
TempoToets Rekenen (TTR) ... 7
Synonym test ... 7
Dutch Auditory & Image Vocabulary Test (DAIVT) ... 8
Peabody Picture Vocabulary Test-III-NL (PPVT) ... 8
Matrix Reasoning Item Bank (MaRs-IB) ... 8
Open Matrix Stimulise Set (OMSS) ... 9
Statistical analysis ... 10
Results ... 10
Descriptive statistics ... 10
T-tests: use of mobile phone or PC ... 14
Correlations between tests ... 15
Reliability ... 16
Mixed models ... 16
Outlier analysis TTR ... 19
Discussion ... 20
Pilot II ... 21
Introduction ... 21
Methods ... 21
Participants ... 21
Procedure ... 21
Test descriptions ... 22
MARQ ... 22

TempoToets Rekenen (TTR) 22
Statistical Analysis 22
Results 23
Descriptive statistics 23
TTR 25
MARQ 28
Discussion 30
General Conclusion 32
References 33
Appendix 35
TTR – version 1 (online home) 35
TTR - version 2 (online lab) 37
TTR – version 3 (pen/paper lab) 39
MARQ – Finnish, English and Dutch translations 41

Foreword

This open-access working report shares preliminary methods and findings in a timely, transparent manner. It is not intended for journal publication. Its primary purpose was to prepare for the later TwinWise data collection by testing procedures, instruments, feasibility, and logistics, so the subsequent large-scale remote study can proceed more efficiently and with higher quality. As a working document, content may be updated as new insights emerge.

Acknowledgements

This work is funded by the European Union (ERC, Project “InterGen”, PI: van Bergen, 101076726), the Research Council of Norway (GenEd grant 335634), and the Dutch Research Council (VIDI Talent Grant, PI: van Bergen, VI.Vidi.221G.007. EvB is a Jacobs Foundation Research Fellow.

Abstract

Background

Online, at-home testing of scholastic skills can widen participation in research, but needs evidence on quality and practicality. We ran two pilots in samples of Dutch adults to: (a) compare a digital version to the official pen & paper version on a short arithmetic test (TempoToets Rekenen; TTR), (b) evaluate two online vocabulary tests, (c) benchmark a developing abstract-reasoning test (OMSS), and (d) examine the Dutch translation of the Math Ability Rating Questionnaire (MARQ) self-report.

Methods

Pilot I (N = 55) included home online testing and supervised lab testing (online and pen/paper). Home: TTR, Matrix Reasoning Item Bank (MaRs-IB), synonym vocabulary test, Dutch Auditory & Image Vocabulary Test (DAIVT). Lab: TTR, synonym test, DAIVT, the Peabody Picture Vocabulary Test (PPVT-III) and the OMSS. Pilot II (N=61) repeated TTR across three settings; home-digital, lab-digital, lab-pen/paper and compared it to (subsets of) the MARQ. Outcomes included internal consistency (α), Pearson correlations (r), intraclass correlation coefficients (ICC) from mixed models, administration time, and feasibility indicators.

Results

Arithmetic (TTR)**:** In Pilot I, cross-format associations were moderate (r = .63). In Pilot II, home-digital vs pen/paper r = .69, lab-digital vs pen/paper r = .70, and home- vs lab-digital r = .79. Scores were generally higher on pen/paper than on the digital versions (modality effect), but the two digital settings were more consistent.

Vocabulary: Both tests were reliable and convergent with PPVT: DAIVT α = .90; r = .76; synonym α = .85; r = .66. The synonym test was shorter than the DAIVT (~4.5 vs. 7 min) and required no audio, favouring remote feasibility; DAIVT provided stronger convergence, likely reflecting the similar format to the PPVT.

Abstract reasoning: OMSS showed acceptable preliminary indices versus MaRs-IB (α = .67; r = .56; ICC = .51; ~11 min). A ~20-item OMSS form appears efficient with limited expected loss in precision.

Self-reported math (MARQ): Internal consistency was strong (α > .80 for MARQ-19/5/3). Higher MARQ scores (poorer self-reported arithmetic ability) were weakly to, at best, moderately correlated with TTR (timed arithmetic performance); the strongest being for MARQ-5 (r = −.33, p = .011).

Conclusions

Remote testing of scholastic skills in adults is feasible. For the remote, online version of the TTR standardized administration (clear on-screen instructions, short practice, consistent device) is crucial to limit modality-related variance. For vocabulary, both online tests are suitable; the synonym test is pragmatic for remote use, whereas DAIVT offers stronger convergence at greater time cost and requires audio. OMSS is promising pending item refinement and shortening. MARQ–TTR associations were weaker than expected, suggesting differences in construct coverage (broad self-report vs. timed operations), translation and sample effects. Within MARQ, the MARQ5 subset appears most suitable in our context due to its brevity, high internal consistency, and the strongest, albeit modest, correlation with TTR.

Pilot I

Introduction

This report documents the findings of two pilot studies on several arithmetic, language, and abstract reasoning tests conducted in preparation for an upcoming large-scale remote data collection. In this project, participants will complete online assessments in their home environment. Therefore, short, reliable, valid online tests in arithmetic and language were needed for adults. The pilot studies reported here focused only on tests intended for adults. The purpose of the first pilot study was to determine the reliability and validity of online arithmetic, vocabulary, and abstract reasoning tests for adults. This report describes measures of reliability (i.e., Cronbach's alpha), validity (i.e., comparison with a gold-standard test), the online application of the tests compared to a version in the lab, and the duration of each test.

Methods

Participants

The sample of the first pilot consisted of 55 Dutch native speakers aged 18 years or older. The participants were primarily first-year bachelor students (Psychology or Pedagogy) from the Vrije Universiteit (VU) Amsterdam, who were gathered through the university participation pool. In addition, participants were gathered from the researchers' social circles. A power analysis was conducted with the program GPOWER (Erdfelder et al., 1996). With 44 participants, there would be 80% power to detect a correlation of .4 (with α set to .05). In addition to the test data, we gathered demographic data of the participants on age, gender, and colour-blindness.

Procedure

The first pilot study consisted of two parts. First, participants took online tests at home (henceforth, home tests). Second, the participants came to the lab to take some online and pen/paper tests (henceforth, lab tests). Table 1 presents an overview of the home and lab tests. All online tests (both at home and in the lab) were set up in Qualtrics, which is an environment for creating online questionnaire using a server hosted by the VU. Participants completed the home tests before coming to the lab. We asked the participants to take the home tests on their mobile phone. While most of the participants performed the tests on their mobile phone (N = 32), several participants used their laptop or PC instead (N = 23). In the lab, the participants first took the pen/paper tests under supervision of a researcher, (in the order of: TempoToets Rekenen (TTR), Peabody Picture Vocabulary Test (PPVT)). Lastly, the online lab tests were taken by the participants on a computer and headphones provided by the lab (in the order of: synonym vocabulary test, Dutch Auditory & Image Vocabulary Test (DAIVT), matrix reasoning test (OMSS)).

Five participants did not come to the lab, so we only have data from the home tests of these participants. Two participants had incomplete data for the arithmetic test in the lab due to disturbances during the test. One participant's audio was not working during their home test, so we excluded their data on the DAIVT which uses audio. Thus, there was complete data from all tests for 47 participants.

Table 1. Overview of tests in the home and lab part of the study.
DAIVT = Dutch Auditory & Image Vocabulary Test. MaRs-IB = Matrix Reasoning Item Bank. OMSS = Open Matrix Stimulise Set, PPVT = Peabody Picture Vocabulary Test.

	Home tests	**Lab tests**	
	Online	**Pen/paper**	**Online**
Arithmetic	Tempo Toets Rekenen	Tempo Toets Rekenen	
Language	Synonym test	PPVT-NL III	Synonym test
	DAIVT		DAIVT
Abstract reasoning	MaRs-IB		OMSS

Test descriptions

TempoToets Rekenen (TTR)

This test of arithmetic fluency was an extended version of the TTR based on the addition/subtraction subtests from the TTR (De Vos, 1992) extended by Elsje van Bergen (2012) to avoid ceiling effects in adults. The test consisted of two parts: 60 addition problems, and 60 subtraction problems of increasing difficulty. For each part, the participant got 60 seconds to correctly solve as many operations as possible, with a short break in between the two parts. This test was included in the online home test, and as a pen/paper test in the lab. It was not included in the online lab test. In the home test, a one-minute timer was set in Qualtrics on each page with operations. In the lab test, the supervising researcher timed the test with their mobile phone. The outcomes of this test were the number of correct answers as the sum score of both parts (max = 120), and the number of correct answers of each part separately.

Synonym test

This vocabulary test (Brysbaert & Vantieghem, 2023) consisted of 40 items, and was conducted as an online test only, both at home and in the lab. In each item, a written word was presented to the participant, together with four written words as answer options. The participant had to choose one out of the four options that best suited the presented written word (see Figure 1 for an example). This could be a synonym, or a term best related to the presented word. There was no time limit for answering these questions. Participants were obligated to answer each item. The outcomes of this test were the number of correct answers, and the total time spent taking the test. This test has been conducted before in a sample of Dutch-speaking adults, where a correlation of .60 was found with the Dutch Author Recognition test, and a Cronbach's alpha of .89 (Brysbaert & Vantieghem, 2023).

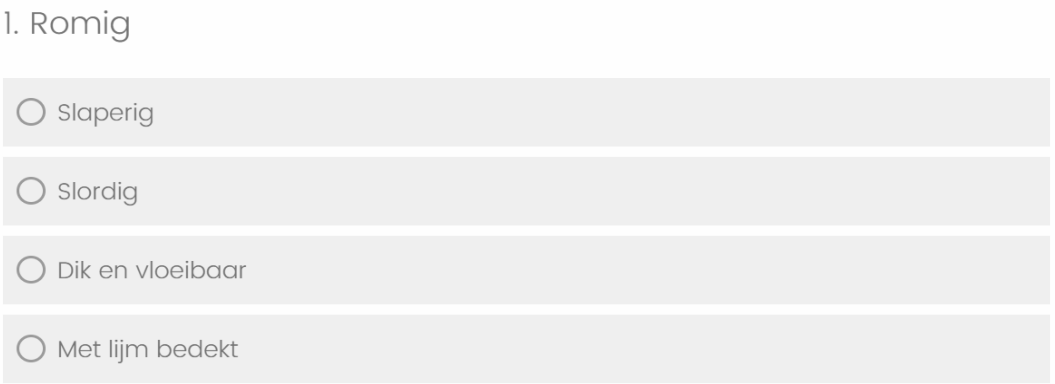


Figure 1. Example question of the vocabulary synonym test.

Dutch Auditory & Image Vocabulary Test (DAIVT)

This vocabulary test (Bousard & Brysbaert, 2021) consisted of 90 items, and was conducted both at home and in the lab. For each item, a word was spoken out loud from a Dutch voice recording, and four pictures were presented as answer options. The participant had to choose the picture that suited the spoken word best. The items can be found on the Open Science Framework: https://osf.io/8kxz7/ (Bousard & Brysbaert, 2020). First, participants got two practice questions: one to check the sound and one to check if they understood the assignment. This test had no time limit, and participants were obligated to answer each item. The outcomes of this test were the number of correct answers and the total time of taking the test. In a previous study, this test had a correlation of .77 with the PPVT-III-NL (Bousard & Brysbaert, 2021)

Peabody Picture Vocabulary Test-III-NL (PPVT)

The Peabody Picture Vocabulary Test (PPVT) was included in this pilot study as a "golden standard" to compare the other vocabulary tests with and was therefore taken only in the lab as pen/paper test under supervision of a researcher. We used the third Dutch version (Schlichting, 2005). The test consisted of multiple sets of 12 items. For each item, the examiner spoke a word out loud, and the participant was presented with four drawings. The participant had to choose the drawing best suited to the spoken word. First, a practice question was presented by the researcher to the participant. After finishing the test, a raw score was calculated. Based on this raw score and the participant's age, a vocabulary quotient and percentile score was determined according to the manual of the test. Only the raw score was used as the outcome for this test, since this relatively old version (2005) made the norm groups less appropriate, and the calculation of the vocabulary quotient and percentile scores were not representative anymore.

Matrix Reasoning Item Bank (MaRs-IB)

This matrix reasoning test consisted of 40 items and was conducted only online in the home test. The original open-access test consists of 80 items (Chierchia et al., 2019), but we reduced this test by selecting 40 items randomly. The items can be found on the Open Science Framework: https://osf.io/g96f4/ (Chierchia et al., 2018). We used items from item set 1; test form 1. We used the non-colourblind friendly version of the items. Half of the items had the minimal difference distractor, and the other half had the paired difference distractor.

For each item, a matrix was present with nine planes: eight of them were filled with a figure, one of them was blank (see Figure 2 for an example). The participant had to choose from four options the figure that could best fill the blank space in the matrix. A practice item was given to the participant first. Participants got 30 seconds for each item. The outcomes of this test was the number of correct responses, the total time spent taking this test, and the time spent on each item.

Figure 2. Example item from MaRs-IB matrix reasoning test. The correct answer is option 4.

Open Matrix Stimulise Set (OMSS)

This matrix reasoning test also consisted of 40 items, randomly selected from a set of 60 items. For each item, a matrix was present with nine planes: eight of them were filled with a figure, one of them was blank (see Figure 3 for an example). The participant had to choose from four options the figure that could best fill the blank space in the matrix. A practice item was given to the participant first. The participant got 30 seconds for each item. The outcomes of this test was the number of correct responses, the total time spent taking this test, and the time spent on each item. For more information on the OMSS see (Kievit, 2023)

The MaRs-IB test was included in this pilot study as “golden standard” to compare the OMSS test with, as the MaRs-IB test is only suitable for adults, whereas the OMSS test will be tested on children as well. Ideally, the same matrix reasoning test will be given to parents and children, as long as the test is reliable and valid in both samples. Since the OMSS was intended for this purpose, it was important to test its reliability and validity by comparing it to another, previously validated test (MaRs-IB).

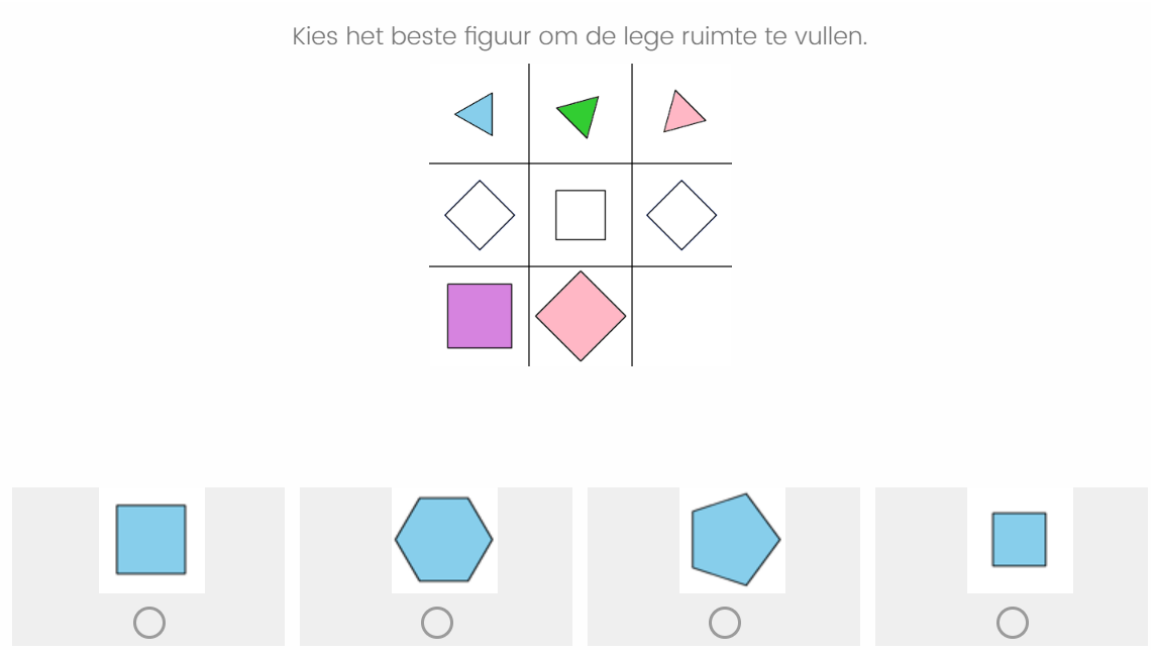


Figure 3. Example item of the OMSS test. The correct answer is option 1.

Statistical analysis

The statistical analyses were performed in R version 4.4.1. (R Core Team, 2024). Descriptive statistics were calculated for the demographic data, the total test scores, and the time spent on each test. T-tests were performed to test if there was a difference in test scores on the TTR taken at home between participants using their mobile phone and participants using their PC. Pearson's product moment correlations were computed between the home and lab tests, and the tests with their golden standards. Lastly, mixed models were fitted to the data to account for correlations due to longitudinal datapoints. The place of taking the test (i.e., home or the lab) was included as a fixed factor, and the participant ID was included as a random intercept. The intraclass correlation (ICC) was computed for each mixed model, representing the proportion of within-participant variance: the higher the ICC, the higher the correlation between the home and lab versions of a test. These mixed models were also applied to comparing the vocabulary tests with the PPVT. In this case, the ICC could be interpreted as the correlation between a vocabulary test and the PPVT, accounted for dependencies in the data. The significance level was set at .05.

Results

Descriptive statistics

The distributions of demographic data are presented in Figure 4. All participants were young adults between the ages of 18 and 27 years old, and were primarily university students. 60% of the sample was female. One participant was colour blind, which was taken into account when analysing the results of the matrix reasoning tests. Although participants were instructed to use their mobile phone in the home test, many used their PC instead (42%). Descriptive statistics of the tests and time spent on each test is presented in Table 2. These descriptive statistics indicated no ceiling effects except for the MaRs-IB, where a ceiling effect is defined as there being no participants scoring higher than twice the standard deviation above the mean. However, the distributions of the test scores do not indicate a ceiling effect for any of the test scores (Figure 5, Figure 6).

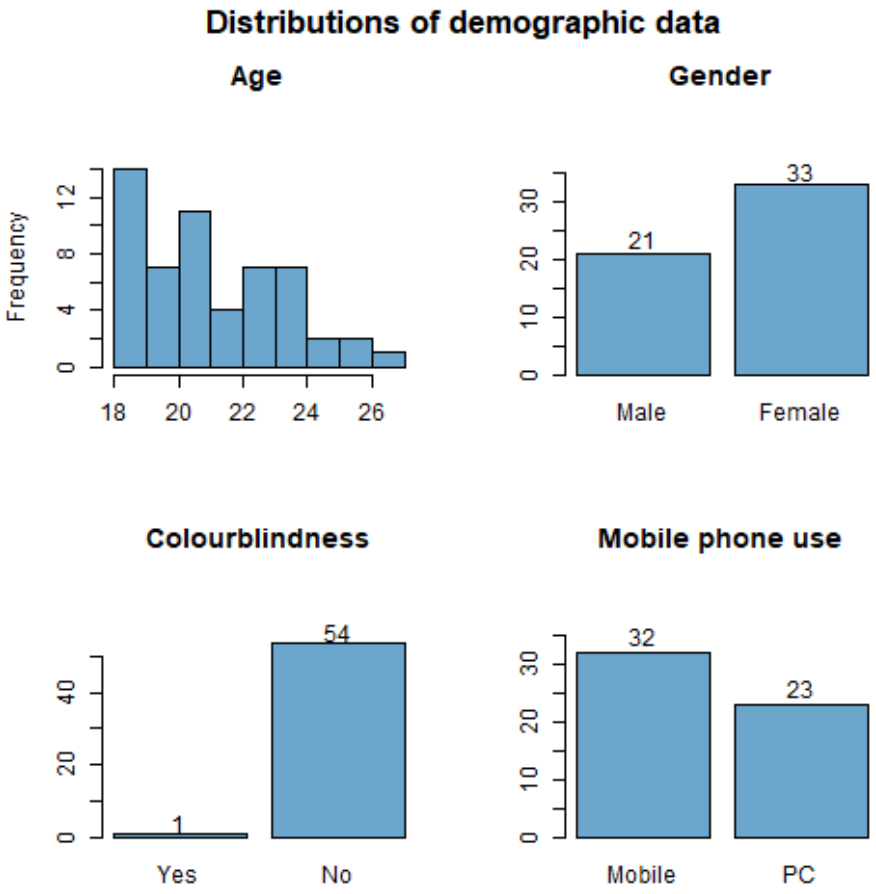


Figure 4. The distribution of age, gender, colour blindness and mobile phone use during the home test of the sample

Table 2. Descriptive statistics of participants' age, test scores, time spent on each test (in minutes), and the time interval between home test and lab test (in days).

	N	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis	SE
Age	55	21.4	2.35	21	18	27	9	0.35	-0.79	0.32
Score TTR - Addition (home)	55	28.04	5.31	29	4	36	32	-2.28	6.75	0.72
Score TTR - Addition (lab)	49	34.73	3.26	34	28	44	16	0.55	0.57	0.47
Score TTR - Subtraction (home)	55	27.36	5.48	28	7	38	31	-1.6	4.02	0.74
Score TTR - Subtraction (lab)	48	32.71	3.18	32	27	42	15	0.46	-0.03	0.46
Score TTR - total (home)	55	55.4	10.49	57	11	73	62	-2.07	5.72	1.41
Score TTR - total (lab)	48	67.52	6.09	67	55	86	31	0.6	0.65	0.88
Score PPVT	50	172.48	10.01	174	146	191	45	-0.63	0.09	1.42
Score synonym (home)	55	21.98	6.91	22	10	38	28	0.21	-0.77	0.93
Score synonym (lab)	50	22.2	7.43	23	7	38	31	0.11	-0.73	1.05
Score DAIVT (home)	54	56.8	12.51	56	33	86	53	0.17	-0.62	1.7
Score DAIVT (lab)	50	58.4	12.65	60	35	85	50	0.1	-0.77	1.79
Score MaRs-IB score (home)	55	26.93	5.59	27	12	37	25	-0.65	0.11	0.75
Score OMSS (lab)	50	28	3.85	29	17	34	17	-0.94	0.94	0.55
Time between home and lab (days)	50	2.72	2.6	1.56	0.11	11.51	11.4	1.29	1.03	0.37
Time synonym test (home)	55	6.58	3.08	5.83	3.13	18.43	15.3	2.09	5.09	0.42
Time synonym test (lab)	50	4.46	1.2	4.32	2.62	8.75	6.13	1.19	2.31	0.17
Time DAIVT (home)	55	13.46	11.93	11.13	6.55	81.35	74.8	4.45	20.56	1.61
Time DAIVT (lab)	50	7.2	1.4	7.09	4.57	11.25	6.68	0.62	0.61	0.2
Time MaRs-IB (home)	55	11.83	3.41	11.77	4.73	26.92	22.18	1.21	5.43	0.46
Time OMSS (lab)	50	11	2.44	10.95	4.2	16.07	11.87	-0.39	0.15	0.35

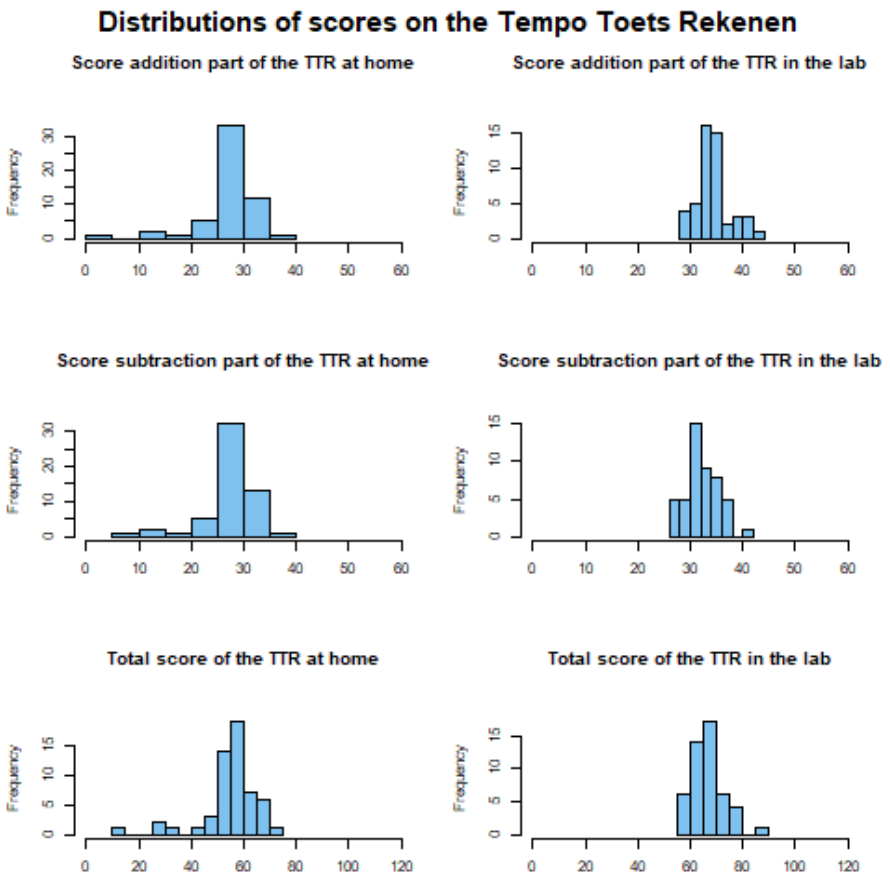


Figure 5. Distributions of scores on the TTR, where scores are calculated as the number of correct answers.

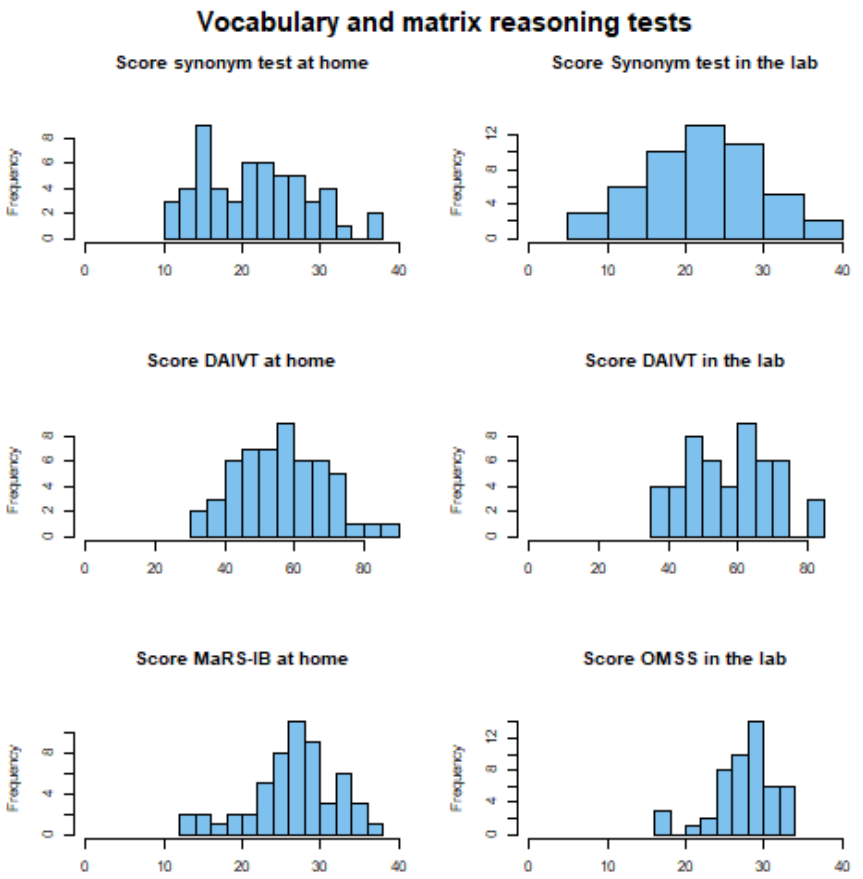


Figure 6. Distributions of test scores on the vocabulary and matrix reasoning tests.

Some participants got very low scores on the TTR at home, while there were no such low scores on the TTR in the lab (Figure 5). Presumably, these participants encountered technical issues while taking the test at home on their phone or PC. Although participants were instructed to not take breaks during the tests, some did (on the DAIVT and the matrix reasoning test; Figure 7). Therefore, the times spent on the tests in the lab were more representative. On average, participants spent 4.5 minutes on the synonym test, 7 minutes on the DAIVT, and 11 minutes on the matrix reasoning test.

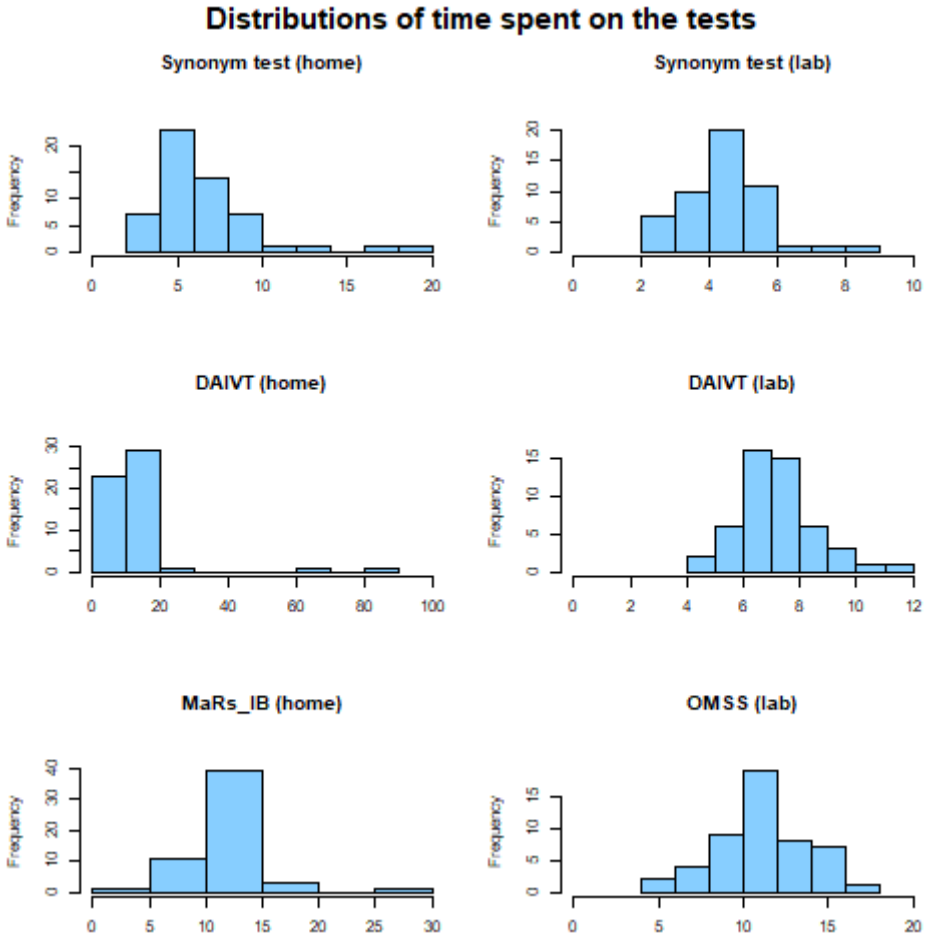


Figure 7. The time in minutes participants spent on each vocabulary and matrix reasoning test.

T-tests: use of mobile phone or PC

Although participants were instructed to use their mobile phone for the home test, more than 40% used their PC instead. Because the answering the questions of the online TTR is very different on phones and computers and because the test is timed, the use of mobile phone or PC may have affected the results. There were no differences in performance due to mobile phone or PC use expected for the other tests. Therefore, we have conducted t-tests for the scores on the TTR where the group of mobile phone users was compared to the group of PC users. Although there was a significant difference for the addition operations and the total score, there was no significant difference between the group of mobile phone users and PC users in the subtraction exercises (Table 3).

Table 3. Results of the t-tests comparing the TTR scores from the home test in the group of mobile phone users to the group of PC users. The null hypothesis of a difference in test score equal to 0 was rejected for the first part of the TTR (addition) and the total score of the TTR. The 95% confidence interval (CI) represents the interval of the difference between the two groups.

	Mean phone users	Mean PC users	t	df	p	95% CI
Addition	29.78	25.61	2.75	27.79	.01	1.07 – 7.28
Subtraction	28.69	25.52	1.98	30.02	.06	-0.10 – 6.44
Total score	58.47	51.13	2.42	28.22	.02	1.12 – 13.56

Correlations between tests

Table 4 shows the correlations between the tests taken at home and in the lab. The arithmetic test showed correlations around .60 between the online home version and the pen/paper version in the lab. Considering both versions contained the exact same test, this correlation was lower than expected and probably due to that there were differences in performance dependent on whether the test was made on pen/paper or digitally. The vocabulary synonym test and DAIVT both showed high correlations between the home and lab versions ($r_{synonym}$ = .93, r_{DAIVT} = .96), which was expected since both were conducted online. The matrix reasoning tests showed a correlation of .56 between the MaRs-IB (home) and the OMSS (lab). Both were conducted online with the same structure, but the tests contained different items.

Table 5 shows the correlations between the vocabulary tests and the golden standard (PPVT). The synonym tests had correlations of .66 and .67 for the home and lab version, respectively. The correlations of the DAIVT were higher, .76 and .79 respectively. The DAIVT is more similar to the PPVT, because the tests are set up the same way: a participant hears a word spoken out loud, and chooses the most appropriate picture or figure. The synonym test is different in structure, which could have resulted in a lower correlation with the PPVT.

Table 4. Pearson's product-moment correlations of test scores taken at home and in the lab.
**Colourblind participant excluded.*

	Correlation	95% CI	t	df	p
TTR – addition	.54	.31 – .72	4.44	47	<.001
TTR – subtraction	.65	.45 – .79	5.80	46	<.001
TTR – total	.63	.42 – .78	5.52	46	<.001
Synonym	.93	.86 - .96	17.21	48	<.001
DAIVT	.96	.94 - .98	24.85	47	<.001
MaRs – OMSS	.56	.33 - .72	4.62	48	<.001
MaRs – OMSS*	.55	.32 - .72	4.49	47	<.001

Table 5. Pearson's product-moment correlations between the vocabulary tests and the raw score of the PPVT.

		Correlation	95% CI	t	df	p
Synonym	**Home**	.66	.46 - .79	6.01	48	<.001
	Lab	.67	.48 - .80	6.20	48	<.001
DAIVT	**Home**	.76	.61 - .86	7.96	47	<.001
	Lab	.79	.66 - .88	9.02	48	<.001

Reliability

Table 6 contains the reliability estimates (Cronbach's alpha) of the vocabulary tests and the matrix reasoning tests. All tests showed high reliability estimates ($\alpha > .80$), except the OMSS ($\alpha = .67$).

Table 6. Reliability coefficient estimates of the vocabulary and matrix reasoning tests.

		Cronbach's alpha	95% CI
Synonym test	**Home**	.85	.79 – .90
	Lab	.87	.82 – .92
DAIVT	**Home**	.90	.86 – .94
	Lab	.91	.87 – .94
Matrix Reasoning test	**MaRs-IB**	.82	.75 – .88
	OMSS	.67	.52 – .79

Mixed models

For mixed models, the test scores of the home and lab versions of the tests were combined together (Figure 8), with the condition specifying whether the test score was from the home test or the lab test (Figure 9). The spaghetti plots in Figure 9 show the test score from a home test on the left side of a graph, and the test score of the same participant from a lab test on the right side, connected by a line to show the "longitudinal" relationship between the two. If the home version is as reliable as the lab version of a test score, it is expected that participants generally show the same pattern or slope in the spaghetti plot, and that lines generally do not cross each other. Since we are interested in measuring differences between participants, it does not matter if all participants score higher (or lower) on the home test compared to the lab test. However, it would be problematic if some participants score higher on the home version, whereas other participants score lower on the home version of a test, changing the rank order of participants.

Participants generally score higher on the pen/paper (lab) version of the TTR, though the rank order was largely preserved. The participants scored fairly similar on the home and lab versions of both vocabulary tests, but they scored quite differently on the matrix reasoning tests. There is no general pattern followed, some participants scored substantially higher in the lab, whereas other scored substantially lower. Moreover, many lines cross each other, indicating a change in rank order.

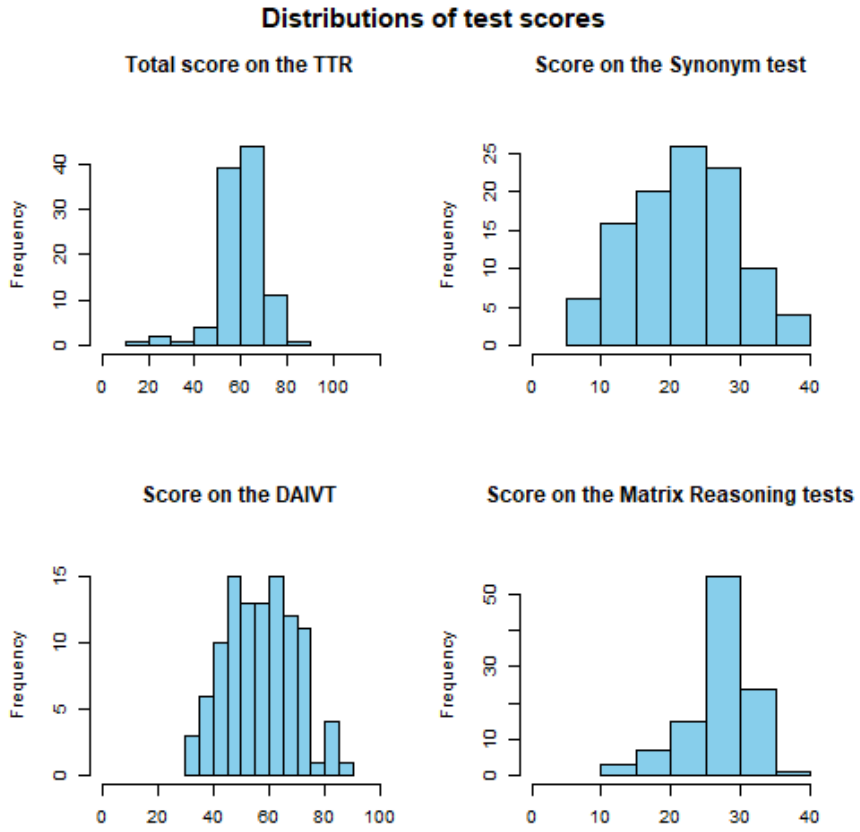


Figure 8. Distributions of test scores with home and lab versions of the tests together. A distribution thus contains two test scores per participant (assuming participants completed the whole test procedure).

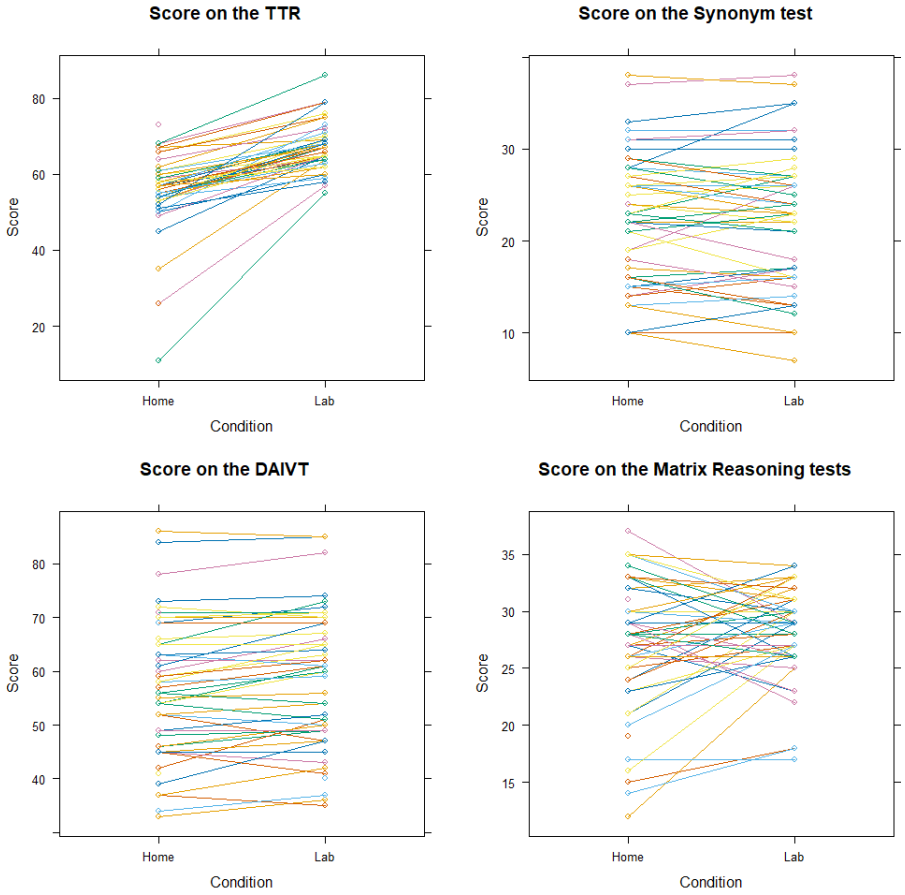


Figure 9. Spaghetti plots of the test scores at home and in the lab.

Table 7 presents the results of the mixed models with Condition (home or lab) as fixed factor, so that the ICC can be interpreted as the correlation between the home and lab versions of a test while accounting for clustering. The ICC in the vocabulary tests are high (.93 for the synonym test, .97 for the DAIVT), making the home versions of both tests proper replacements of the lab versions. However, the ICCs were lower for the TTR and matrix reasoning tests (.60 and .51, respectively). The matrix reasoning tests contained different items, but were set up in the same way. The TTR had the exact same items, but the difference was in a digital version at home and a pen/paper version in the lab.

Table 7. Results of mixed models with Condition (home or lab) as fixed factor and the participants as random intercept. The intercept is the mean test score on the home version of the test, and Beta is the difference with the test score of the lab version. The intraclass correlation (ICC) was computed from the models to determine to what extent the rank order was preserved, and can be interpreted as the correlation between the home and lab versions of the tests, accounting for clustering within participants.

	Intercept (SE)	**Beta (SE)**	**ICC**
TTR	55.40 (1.19)	12.06 (1.13)	.60
Synonym test	21.98 (0.96)	0.15 (0.39)	.93
DAIVT	56.47 (1.71)	1.79 (0.48)	.97
Matrix Reasoning	26.93 (0.65)	1.03 (0.67)	.51

Because a large proportion of participants used their PC instead of their mobile phone, we hypothesized that there may have been differences between the mobile phone users and PC users in how well the pen/paper version of the TTR translates to a digital versions. Additionally, there may have been differences in images and sound between mobile phones and PCs on the DAIVT. Therefore, the mixed models were fitted again for the TTR and the DAIVT on subsamples of mobile phone users and PC users (Table 8). The DAIVT showed fairly similar results to the full sample, with high ICCs in the subsample. However, the TTR showed substantial differences in ICC between mobile phone users and PC users. The correlation between the digital home version of the test and the pen/paper version at the lab was much higher in mobile phone users (.78) than in PC users (.52).

Table 8. Results of mixed models of TTR and DAIVT test scores, with Condition (home or lab) as fixed factor, split into samples of mobile phone users and PC users. The intercept is the mean test score on the home version of the test, and Beta is the difference with the test score of the lab version. The intraclass correlation (ICC) was computed from the models to determine to what extent the rank order was preserved, and can be interpreted as the correlation between the home and lab versions of the tests, accounting for clustering within participants.

		Intercept (SE)	**Beta (SE)**	**ICC**
TTR	**Mobile**	58.47 (1.11)	9.86 (0.80)	.78
	PC	51.13 (2.24)	15.18 (2.28)	.52
DAIVT	**Mobile**	57.66 (1.88)	2.00 (0.69)	.94
	PC	54.83 (3.14)	1.45 (0.64)	.98

Table 9 shows the results of mixed models of the vocabulary test scores compared to the golden standard. The ICC can be interpreted as the correlation between a vocabulary test and the golden standard, accounting for clustering. The ICCs were a bit lower than the correlations (Table 5), but fairly similar. The ICCs of the DAIVT were higher than the ICCs of the synonym tests, both the home and lab versions. For both tests, the ICC was higher in the lab version than the home version, but the differences were small.

Table 9. Results of mixed models of vocabulary test scores, comparing the tests with the golden standard (PPVT). The intraclass correlation (ICC) was computed from the models to determine to what extent the rank order was preserved, and can be interpreted as the correlation between the vocabulary test and the golden standard, accounting for clustering within participants.

		ICC
Synonym	**Home**	.61
	Lab	.64
DAIVT	**Home**	.74
	Lab	.77

Outlier analysis TTR

Regarding the TTR, we noticed some participants scored really low on the digital home test (e.g., a sub-score of 4 out of 60 answers correct), while scoring much higher on the pen/paper version in the lab. Some participants could have gotten technical issues with the home version of the test, resulting in very low scores. Therefore, this outlier analysis performed the same statistical analyses as before, but with removal of participants having a total score on the TTR lower than 3 standard deviations below the mean.

One participant was excluded, having a total score on the TTR of 11 on the home test, and 55 on the lab test. Participants using the PC for the home test tended to score lower than participants using their mobile phone (Mean mobile = 58.47, Mean PC = 52.95, t = 2.19, df = 30.18, p = .04). The correlation between the home and lab version of the total scores was .60 (95% CI = .38 – .76), which was lower than the correlation found without removal of outliers. The results of the mixed models are presented in Table 10, showing a slightly higher ICC now the outlier was removed (.60 vs. .65). The ICC in mobile phone users remained unchanged, but it slightly increased in PC users (.52 vs. .54), which was expected since the outlying participant used their PC to conduct the home test.

Table 10. Results of mixed models of the TTR with Condition (home or lab) as fixed factor and the participants as random intercept, with removal of participants scoring lower than 3 standard deviations below the mean test score. The models were fitted on three (sub)samples: the full sample, the subsample of mobile phone users, and the subsample of PC users.

	Intercept (SE)	Beta (SE)	ICC
Full sample	56.22 (1.05)	11.42 (0.93)	.65
Mobile	58.47 (1.11)	9.85 (0.80)	.78
PC	52.96 (1.87)	13.81 (1.87)	.54

Discussion

This report describes the first pilot study that investigated the reliability and validity of online tests on arithmetic ability, vocabulary, and abstract reasoning aimed at Dutch adults. Regarding arithmetic ability, we aimed to answer the question whether the digital version of the TTR measures arithmetic ability as well as the original pen/paper version. The total scores of the two versions had a Pearson's product-moment correlation of .63 (95% CI: .42 – .78), and the ICC computed from the mixed model was .60, which can be interpreted as a within-person correlation between the two test scores. These results were lower than expected, which could have been caused by almost half of participants conducting the digital version of the test on a PC, rather than on their mobile phone as instructed. Moreover, we did not give explicit instructions on how to enter the answers on their phone (or PC) and how to get to each next question. These two factors could have created variability between the participants' scores, which did not reflect their arithmetic ability. Therefore, this part was piloted again in the second pilot study, with clear instructions, practice items, and consistent use of one device across participants, to prevent variability created by technicalities.

For vocabulary, our goal was to determine which test would be most appropriate in an online, remote setting, based on internal consistency (Cronbach's alpha), average testing time, and correlation with a gold-standard measure (PPV). Although the DAIVT showed higher reliability (α = .90) and stronger associations with the PPVT (r = .76; ICC = .74), the synonym test also performed well (α = .85; r = .66; ICC = .61) with only a modestly lower correlation. Participants completed the synonym test more quickly (≈4.5 minutes vs. 7 minutes for the DAIVT), and it does not require audio, improving feasibility in remote contexts. Moreover, a somewhat lower correlation with the PPVT was expected for the synonym test because, unlike the DAIVT, it does not share the PPVT's format (spoken word with four picture options), reducing shared method variance.

Regarding abstract reasoning, we aimed to answer the question whether the OMSS measured abstract reasoning as well as the MaRs-IB. Although the MaRs-IB is an open-source, validated item bank for matrix reasoning tests, the OMSS was of interest because it will include items appropriate for children of different ages. These were still being tested in another lab (Kievit, 2023), and we have therefore used rough items that had not been analyzed yet. The reliability of the OMSS was .67, and the Pearson's correlation and ICC with MaRs-IB were .56 and .51, respectively. On average, participants spent 11 minutes taking the OMSS test. Given that the items would become better and other pilot studies were still ongoing, these results were satisfactory enough to include the OMSS in the final test battery for adults and children. However, the test will be shortened to 20 items, to reduce the test length, as we expect that the test still perform well enough in measuring abstract reasoning.

Pilot II

Introduction

This pilot study was conducted to determine the performance of the online application of the TempoToets Rekenen (TTR) comparing to the original pen and paper version, and to investigate the performance of the Dutch translation of the Math Ability Rating Questionnaire (MARQ; Khanolainen et al., 2025; Khanolainen, Van Bergen, Tolvanen, et al., 2023). The TTR was investigated already in the previous pilot study in adults, but many participants did not adhere to the instructions of taking the test on the mobile phone and used their computer instead. The differential use of mobile phones and computers created variance in the data that could not be taken into account when analysing the results due to a lack of power. Therefore, we investigated the online application of the TTR again in a second pilot study in adults, where all participants were forced to take the test on a PC, to prevent variance created by different devices. With this second pilot, the goal was to answer two main research questions:

1. Investigate the reliability of the digital remote version of the TTR compared to the pen-paper lab setting.
2. Evaluate the Dutch MARQ for mathematics ability in our context and choose the item subset that offers the best balance of coverage and brevity.

To answer the first question, we compared repeated TTR administrations across three settings—home-digital, lab-digital, and pen/paper—using correlations and mixed-effects models to account for within-person dependence from longitudinal repeated measures. The second question is answered by investigating the Cronbach's alpha for each MARQ-item and by comparing the correlations between the paper-pencil TTR, individual MARQ-items and the sum scores of the shortened versions of the MARQ.

Methods

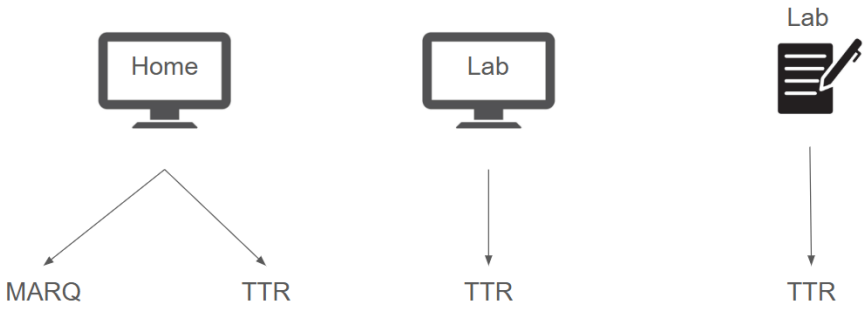
Participants

This study contained 61 participants, gathered through the university participation pool. Except for one, participants were all first-year Psychology or Pedagogy bachelor students. 57 participants had complete data on all tests. In addition to test data, we gathered demographic data of the participants on age and gender.

Procedure

This study consisted of two parts. First, participants conducted the MARQ and performed the online TTR at home on their PC. The test could not be taken on a mobile phone, this option was blocked. Second, participants came to the lab to take the TTR again online at a PC. Lastly, they took the TTR on pen and paper under supervision of a researcher. Thus, the participants took the TTR three times (Figure 1), each a different version. All online tests (both at home and in the lab) were set up in Qualtrics.

Figure 1. Overview of the tests in the home and lab part of the study.
MARQ = Math Ability Rating Questionnaire; TTR = TempoToets Rekenen.



Test descriptions

MARQ

The Math Ability Rating Questionnaire (MARQ) was developed to measure self-reported mathematical skills in adults (Khanolainen, Van Bergen, Koponen, et al., 2023). Initially, the test consisted of 30 items, and was reduced to the 19 items (MARQ-19) with the strongest associations with tested mathematical skills (correlation equal to or greater than .40). In addition, two shorter version of the MARQ (MARQ-3 and MARQ-5) were created with the most discriminating items based on a confirmatory factor analysis and item response theory model (items 2, 3 and 8 for MARQ-3, items 1, 2, 3, 8 and 13 for MARQ-5). All questions had a Likert scale as response ranging from 1 to 5. A high score on the MARQ indicates a low perceived math ability.

We translated the MARQ-19 to Dutch with help of two Dutch-Finnish researchers, and included this in the online home test. The translated version of the MARQ can be found in the appendix. Participants were not forced to answer each question, but were reminded once if they did not answer a question.

TempoToets Rekenen (TTR)

This test of arithmetic fluency was an extended version of the TTR, based on the addition/subtraction subtests from the TTR (De Vos, 1992), extended first by Elsje van Bergen (2012) to avoid ceiling effects in adults. The test consisted of two parts: 60 addition problems, and 60 subtraction problems of increasing difficulty. For each part, the participant got 60 seconds to correctly solve as many operations as possible. The outcomes of these tests were the number of correct answers as the sum score of both parts (max = 120), and the number of correct answers of each part separately. Three different versions were used: two online, and one pen/paper version (see Appendix). In the online versions, a one-minute timer was set in Qualtrics on each page with operations. With the pen/paper test, the supervising researcher timed the test with their phone. Participants first got 10 practice questions in the online versions and were instructed to use the TAB button on the computer to go to each next question.

Statistical Analysis

The statistical analyses were performed in R version 4.4.1. (R Core Team, 2024). Total sum scores were calculated for the three versions of the TTR, as well as total sum scores for the MARQ-19, MARQ-5 and MARQ-3. Items 1, 2, and 30 of the MARQ were reverse coded. Descriptive statistics

were calculated for the demographic data, the total scores of the TTR and MARQ, and the item responses of the MARQ. The total scores were checked for outliers.

Three types of statistical analysis were performed. First, the Cronbach's alpha was computed for each version of the MARQ to investigate the internal consistency of the item set. Second, the Pearson's product moment correlations were computed between three versions of the TTR. Third, correlations were computed between the test score of the pen/paper TTR version and each MARQ item, the MARQ-19, MARQ-5 and MARQ-3 score. Last, mixed models were fitted to the TTR scores to account for correlations due to longitudinal datapoints. The place of taking the test (i.e., home or the lab) was included as a fixed factor, and the participant ID was included as a random intercept. The intraclass correlation (ICC) was computed for each mixed model, representing the proportion of within-participant variance: the higher the ICC, the higher the correlation between the home and lab versions of a test. The significance level was set at .05.

Results

Descriptive statistics

Table 2 presents descriptive statistics on participants' age, their TTR scores and item responses on the MARQ. The distributions of the demographic data are shown in Figure 2 and 3. All participants were young adults between the age of 17 and 26 years old, and were primarily first-year university students. 95% of the participants was female. Despite specific instructions, 14 participants attempted to perform the home-test on a mobile device. However, due to the Qualtrics set-up participants were not able to continue the test on a mobile device and thereby all participants performed the tests on a computer. In Figure 4 and 5 the distributions of the TTR and MARQ scores are represented, respectively.

Table 2. Descriptive statistics of the participants' age in years, their scores on the TTR, and item responses on the MARQ.

	N	Mean	SD	Median	Min	Max	Skew	Kurtosis	SE
Age	60	19.38	1.55	19.00	17.00	26.00	1.71	4.13	0.20
Addition (Home)	60	29.17	3.30	29.50	19.00	35.00	-0.71	0.37	0.43
Addition (Lab)	59	29.34	2.87	30.00	18.00	35.00	-1.24	2.56	0.37
Addition (Pen/paper)	58	35.64	3.43	36.00	27.00	44.00	-0.04	-0.18	0.45
Subtraction (Home)	60	27.53	3.28	28.00	17.00	34.00	-0.35	0.48	0.42
Subtraction (Lab)	59	25.83	3.12	26.00	18.00	32.00	-0.30	-0.08	0.41
Subtraction (Pen/paper)	59	32.20	3.25	33.00	24.00	39.00	-0.23	-0.15	0.42

	N	Mean	SD	Median	Min	Max	Skew	Kurtosis	SE
TTR score (Home)	60	56.70	6.00	57.00	41.00	68.00	-0.52	0.12	0.77
TTR score (Pen/paper)	58	67.90	6.35	68.00	54.00	82.00	-0.11	-0.38	0.83
Item 1 **(MARQ5)**	61	2.43	0.67	2.00	1.00	4.00	-0.66	-0.35	0.09
Item 2 **(MARQ3/5)**	61	2.18	0.92	2.00	1.00	4.00	0.77	-0.18	0.12
Item 3 **(MARQ3/5)**	61	1.93	1.09	2.00	1.00	5.00	1.11	0.41	0.14
Item 4	61	2.21	1.17	2.00	1.00	5.00	0.51	-1.05	0.15
Item 5	61	2.84	0.73	3.00	1.00	5.00	0.50	1.10	0.09
Item 6	61	1.79	0.97	1.00	1.00	5.00	1.08	0.55	0.12
Item 8 **(MARQ3/5)**	61	1.87	1.07	2.00	1.00	5.00	1.37	1.43	0.14
Item 9	61	1.77	0.90	2.00	1.00	5.00	1.26	1.57	0.12
Item 11	61	1.67	0.77	2.00	1.00	4.00	0.83	-0.15	0.10
Item 12	61	2.64	1.13	3.00	1.00	5.00	0.04	-1.00	0.14
Item 13 **(MARQ5)**	61	1.56	0.87	1.00	1.00	5.00	1.72	3.00	0.11
Item 15	61	1.97	0.91	2.00	1.00	5.00	0.84	0.57	0.12
Item 16	61	1.82	0.83	2.00	1.00	4.00	0.86	0.24	0.11
Item 17	61	1.89	0.78	2.00	1.00	4.00	0.40	-0.68	0.10
Item 18	61	2.08	0.97	2.00	1.00	4.00	0.38	-1.01	0.12
Item 23	61	2.28	1.21	2.00	1.00	5.00	0.57	-0.77	0.16
Item 26	61	2.70	1.13	3.00	1.00	5.00	0.18	-0.71	0.14
Item 29	61	2.90	1.11	3.00	1.00	5.00	0.19	-0.59	0.14

	N	Mean	SD	Median	Min	Max	Skew	Kurtosis	SE
Item 30	61	2.46	0.83	2.00	1.00	5.00	0.56	0.33	0.11
MARQ19	61	40.98	9.71	40.00	20.00	63.00	0.22	-0.57	1.24
MARQ3	61	5.98	2.67	5.00	3.00	14.00	1.27	0.93	0.34
MARQ5	61	9.97	3.59	9.00	5.00	22.00	1.38	1.74	0.46

Figure 2.

Distribution of the age and gender of the participant.

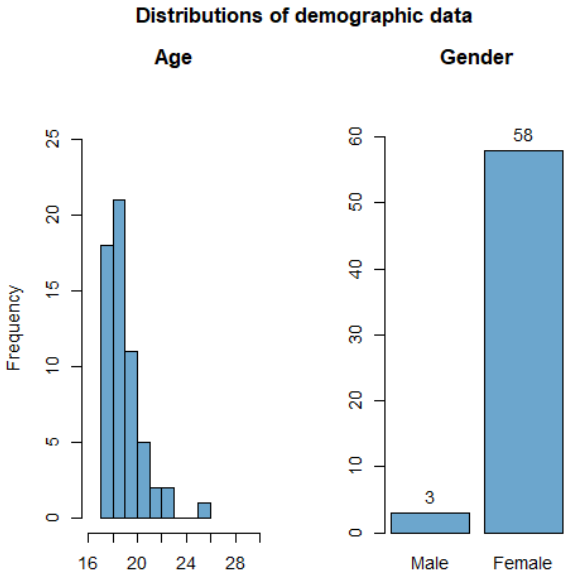
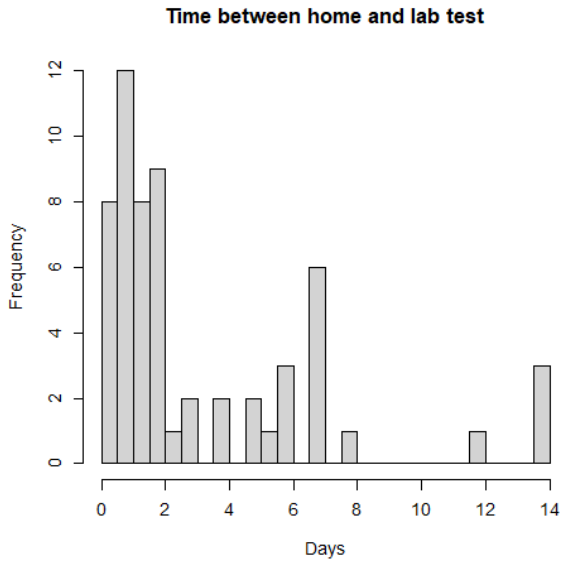


Figure 3.

Distribution of time between the testing moments



TTR

The total TTR score had a mean of 67.90 (SD = 6.35, range: 54–82) in the pen/paper version, 55.17 (SD = 5.49, range: 40–67) in the lab setting, and 56.70 (SD = 6.00, range: 41–68) at home. Correlations between test settings and subtests are presented in Table 3. The correlation between addition and subtraction subtests ranged from .66 (home) to .68 (lab) and .81 (pen/paper). The total TTR score correlated at .79 (95% CI: .66–.87) between home and lab settings on the computer, .69 (95% CI: .52–.81) between home and pen/paper, and .70 (95% CI: .54–.81) between lab and pen/paper. Mixed models incorporating test setting as a fixed effect yielded adjusted intraclass correlation coefficients (ICCs) of .68 for home vs. pen/paper, .69 for lab vs. pen/paper, and .78 for lab vs. home. These results in combination with the correlation analyses, suggest that while all settings demonstrate relatively strong reliability, the digital-lab and digital-home modality are the most consistent with each other.

Table 3. correlations between the different subtests and test settings of the TTR.

	Correlation	Test statistic	Degrees of freedom	P-value	95% Confidence interval
Addition					
Home - Lab	.64	6.18	56	7.782e-08	.45 - .77
Home - Pen/paper	.64	6.22	55	6.955e-08	.46 – .77
Lab - Pen/paper	.57	5.2	56	2.908e-06	.37 – .72
Subtraction					
Home - Lab	.67	6.77	56	8.239e-09	.50 – .79
Home - Pen/paper	.58	5.3	56	2.008e-06	.38 – .73
Lab - Pen/paper	.65	6.53	57	1.9e-08	.48 – .78
Total Score					
Home - Lab	.79	9.50	56	2.884e-13	.66 – .87
Home - Pen/paper	.69	7.06	55	2.998e-09	.52 – .81
Lab - Pen/paper	.70	7.36	56	8.803e-10	.54 – .81
Addition - Subtraction					
Home	.66	6.75	58	7.635e-09	.49 – .79
Lab	.68	6.99	57	3.262e-09	.51 – .80
Pen/paper	.81	10.26	56	1.761e-14	.69 – .88

Note. All correlations were significantly different from 0, tested with an alpha of .05.

Figure 4. Distribution of the scores on the subtests of the TTR.

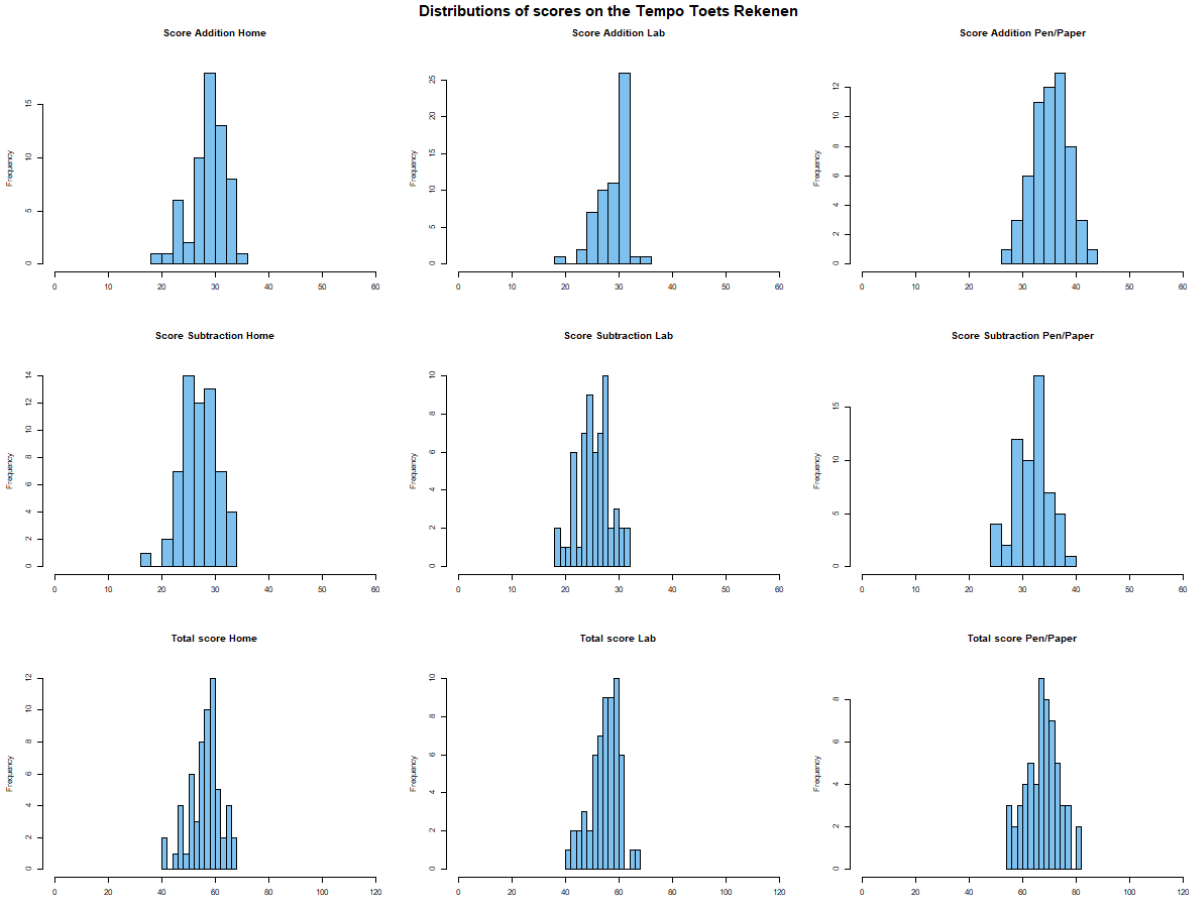
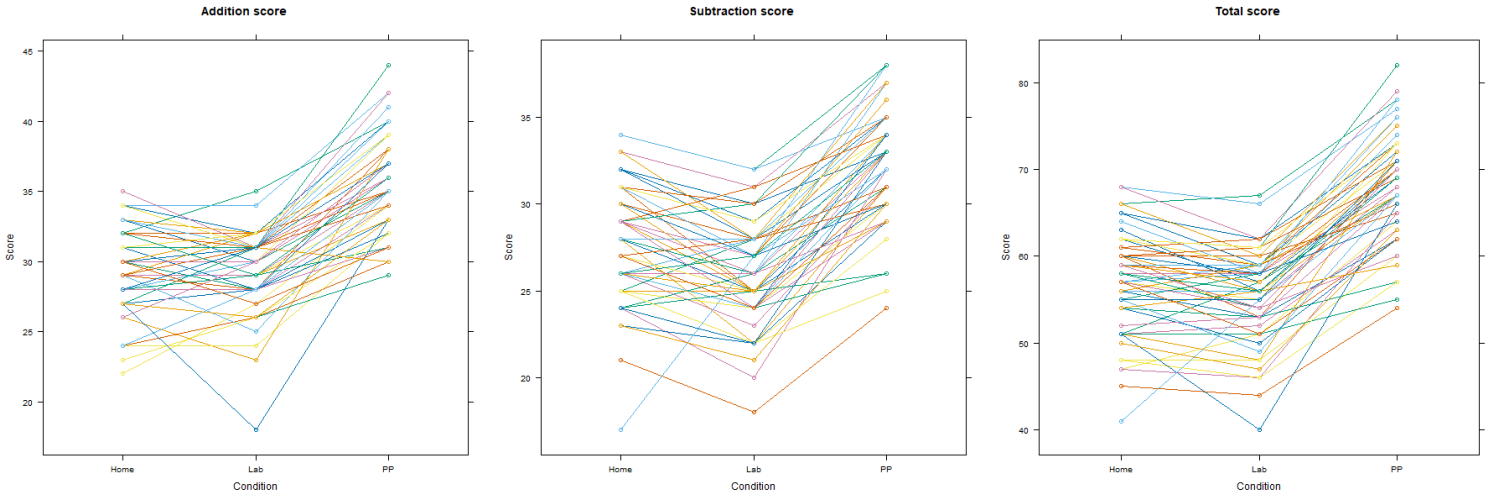


Figure 5. Spaghetti plot of TTR-scores based on different testing setting.



MARQ

The internal consistency of the MARQ subsets was assessed using Cronbach's alpha. The reliability was strong across the subsets, with MARQ19 (α = .86), MARQ3 (α = .83), and MARQ5 (α = .82). The relationships between the MARQ subsets and the total score of the pen/paper TTR were examined using Pearson correlations and tested with an alpha of .05 (Table 4). As expected, the majority of correlations were negative, given the reverse-coded nature of the MARQ items. Among the MARQ subsets, MARQ19 showed a weak negative correlation with the TTR (r = -.23, p = .089) and this correlation was not significantly different form 0. In contrast, MARQ3 demonstrated a statistically significant negative correlation (r = -.29, p = .027). MARQ5 exhibited the strongest, significant negative correlation (r = -.33, p = .011), suggesting that this subset of items is most closely associated with the total score. Several individual items displayed significant negative correlations, further supporting the relationship between MARQ scores and pen/paper TTR performance. The most notable items included Item 1 (r = -.32, p = .016), Item 3 (r = -.29, p = .027), Item 8 (r = -.27, p = .040), and Item 30 (r = -.35, p = .007). When analyzing the three strongest individual items identified in the pilot study (Items 1, 3, and 30), their combined correlation with the total score was even more pronounced (r = -.38, p = .003). Although most correlations were negative, Item 26 (r = .16, p = .218) and Item 12 (r = .01, p = .915) were weakly positive; however, these were not statistically significant different from 0.

Figure 6. Distribution of the scored on the subsets of the MARQ.

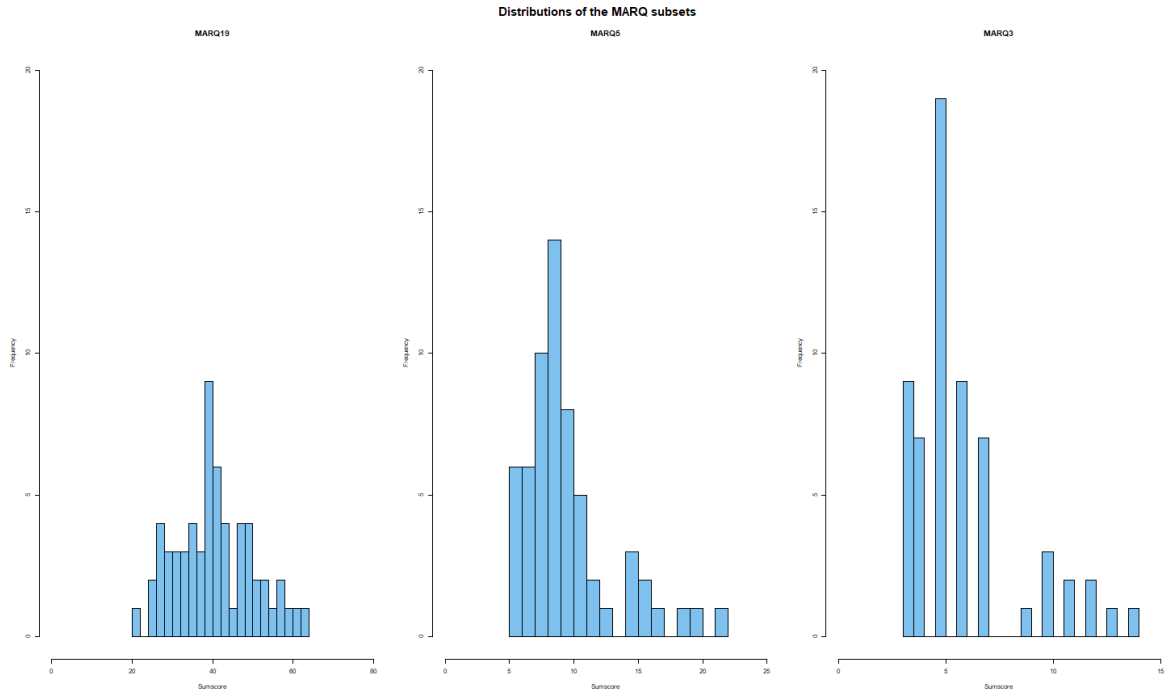


Table 4. Correlations between the MARQ items and the total score of the pen/paper TTR.

	Correlation	T-test	Degrees of freedom	p-value	95% CI
Item 1	-.32	-2.49	56	.016	-.53 – -.06
Item 2	-.17	-1.26	56	.212	-.41 – .09
Item 3	-.29	-2.27	56	.027	-.51 – -.03
Item 4	-.02	-.11	56	.909	-.27 – .24
Item 5	-.18	-1.38	56	.173	-.42 – .08
Item 6	-.23	-1.77	56	.081	-.46 – .03
Item 8	-.27	-2.10	56	.040	-.49 – -.01
Item 9	-.13	-1.01	56	.314	-.38 – .12
Item 11	-.07	-.50	56	.622	-.32 – .20
Item 12	.01	.11	56	.915	-.24 – .27
Item 13	-.20	-1.53	56	.131	-.44 – .06
Item 15	-.24	-1.82	56	.075	-.47 – .02
Item 16	-.05	-.38	56	.699	-.31 – .21
Item 17	-.10	-.76	56	.450	-.35 – .16
Item 18	-.02	-.12	56	.901	-.27 – .24
Item 23	-.01	-.04	56	.965	-.26 – .25
Item 26	.16	1.25	56	.218	-.09 – .41
Item 29	-.07	-.58	56	.565	-.33 – .18
Item 30	-.35	-2.78	56	.007	-.56 – -.10
MARQ19	-.23	-1.73	56	.089	-.46 – .03
MARQ3	-.29	-2.26	56	.027	-.51 – -.03
MARQ5	-.33	-2.62	56	.011	-.54 – -.08

Top 3 from pilot (item 1, 3, 30)	-.38	-3.11	56	.003	-.58 – -.14

Note. Items in **bold** are part of MARQ5, items that are underlined are part of MARQ3.

Discussion

This second pilot study aimed to evaluate the reliability of the digital, remote version of the TempoToets Rekenen (TTR) compared to its traditional pen/paper counterpart. It also assessed the feasibility of a shortened, translated version of the *Math Ability Rating Questionnaire* (MARQ) in a Dutch sample.

The findings indicate that the online version of the TTR demonstrates relatively strong reliability, with the most critical comparison being between the home digital version (the intended setting) and the pen/paper version (the original setting). The correlation between these two settings was moderate (r = .69), suggesting that while the digital version performs well, it does not yet fully replicate the pen/paper format. The correlation between the lab digital and pen/paper versions was similar (r = .70, ICC = .69), while the correlation between the home and lab digital settings was highest (r = .79, ICC = .78). Participants performed better in the pen/paper version than in both digital versions, indicating an effect of test modality. These results suggest that while all versions show reasonable reliability, the digital home and digital lab versions are the most consistent with each other. However, the correlation is high enough to assume that the remote, digital version of the TTR measures the same arithmetic skills as the pen/paper version.

Regarding the MARQ, the translated Dutch version exhibited strong internal consistency across different subsets, with Cronbach's alpha values above .80 for MARQ19, MARQ5, and MARQ3. This suggests that the translated instrument retains the psychometric properties of the original version. However, correlations between MARQ subsets and TTR performance were mostly weak or non-significant, with MARQ-5 showing the strongest association (r = −0.33, p = .011). Nonetheless, these unexpectedly low correlations warrant further investigation into potential translation effects or cultural differences in self-reported mathematical ability.

Several factors may underlie these unexpected patterns. First, some issues may stem from the translation process, where subtle differences in wording could have altered item interpretation. Second, because the sample was Dutch, predominantly female, and comprised very young adults (M = 19.38, SD = 1.55) who were largely psychology and pedagogy students, some items may have shown reduced discriminative power. Most importantly, however, it is likely that the two instruments capture fundamentally different constructs: the TTR taps into a relatively narrow domain of arithmetic skills, focusing primarily on timed, automated operations, whereas the MARQ is designed to assess a broad spectrum of mathematical abilities. Further research is needed to disentangle the relative contribution of these factors and to clarify the extent to which translation issues, sample-specific item functioning, and construct coverage account for the observed associations.

In conclusion, this study provides evidence for the reliability of the digital TTR across different settings but highlights the need for caution when interpreting results from the home digital version, as it may capture variance due to the testing modality rather than mathematical ability. The strong correlation between the home and lab digital versions suggests that digital testing environments can

be highly consistent, even when used remotely. Additionally, MARQ5 showed the strongest association with the TTR, indicating that it may be the most suitable version for use in a Dutch sample. However, the MARQ correlations were not particularly strong, suggesting that further refinement or validation may be needed to ensure the questionnaire accurately reflects self-reported mathematical ability in this population.

General Conclusion

Remote, digital assessment across arithmetic, vocabulary, and reasoning appears feasible, but standardization and instrument choice do matter. For arithmetic (TTR), the first pilot underscored the need for tighter administration to limit non-ability variance, for instance, explicit on-screen instructions, a brief practice block, and a single device type. In the second evaluation, the digital TTR showed relatively strong reliability, with the critical comparison between home digital (intended setting) and pen/paper (original setting) yielding r = .69 (ICC = .68); the lab digital–pen/paper correlation was similar (r = .70; ICC = .69), and home digital–lab digital was highest (r = .79; ICC = .78). Participants scored higher on pen/paper than on both digital versions, evidencing a modality effect. Taken together, all versions were reasonably reliable, the two digital settings were most consistent with each other, and the correlations are high enough to support that the remote, digital TTR taps the same arithmetic skills as the pen/paper test—provided administration is standardized.

For vocabulary, both the synonym test and the DAIVT showed high internal consistency and good convergence with the PPVT. DAIVT performed better (α = .90; r = .76; ICC = .74), which is expected given its shared spoken-word/picture-choice format with PPVT. The synonym test also performed well (α = .85; r = .66; ICC = .61), was shorter (~4.5 vs 7 minutes), and does not require audio — advantages that reduce failure points in remote settings. Given its brevity and robustness, the synonym test is a strong option for remote use, whereas DAIVT provides stronger convergence but takes over twice as long.

For abstract reasoning, preliminary 40-items OMSS results against MaRs-IB were acceptable given ongoing item development (α = .67; r = .56; ICC = .51; ~11 minutes). Moving to a ~20-item form should improve efficiency with limited loss of precision.

Finally, the Math Ability Rating Questionnaire (MARQ; Khanolainen et al., 2025) was translated from Finnish to Dutch and administered in its full 19-item version. Internal consistency was strong (α > .80), indicating adequate reliability. Correlations between the Dutch MARQ and timed arithmetic fluency were generally weak, and most item-level associations were not statistically significant. However, the brief five-item version (MARQ-5; Khanolainen et al., 2025) showed a moderate negative correlation with arithmetic fluency (r = –.33, p = .011). As the MARQ was designed to capture broader self-perceived math ability rather than speeded calculation, we would expect higher correlations with a more comprehensive mathematics test battery—indeed, in the original Finnish validation, the MARQ-5 correlated r = .75 with a broad latent factor of tested mathematical skills (Khanolainen et al., 2025).

This pattern may reflect several factors: (a) conceptual differences between constructs (TTR assesses narrow, timed arithmetic fluency, whereas MARQ captures broader self-perceived math difficulties), (b) translation nuances or cultural differences affecting item interpretation, and (c) restricted variance in this relatively young, high-performing university sample. Despite these limitations, the Dutch MARQ demonstrated excellent internal consistency, and the MARQ-5 short form—comprising five items identified by Khanolainen et al. (2025) as the most informative—showed the highest convergence with objective arithmetic performance. The MARQ-5 thus appears most suitable for inclusion in upcoming TwinWise data collections, while future research with larger and more diverse Dutch samples should further examine its validity and cross-linguistic comparability.

References

Bousard, I., & Brysbaert, M. (2020). *Dutch Auditory & Image Vocabulary Test (DAIVT)*. https://doi.org/10.17605/OSF.IO/8KXZ7

Bousard, I., & Brysbaert, M. (2021). The Dutch Auditory & Image Vocabulary Test (DAIVT): A New Dutch Receptive Vocabulary Test for Students. *Psychologica Belgica*, *61*(1), 1. https://doi.org/10.5334/pb.552

Brysbaert, M., & Vantieghem, A. (2023). No Correlation Between Articulation Speed and Silent Reading Rate when Adults Read Short Texts. *Psychologica Belgica*, *63*(1), 82–91. https://doi.org/10.5334/pb.1189

Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S.-J. (2019). The matrix reasoning item bank (MaRs-IB): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society Open Science*, *6*(10), 190232. https://doi.org/10.1098/rsos.190232

Chierchia, G., Fuhrmann, D., Knoll, L., Piera Pi-Sunyer, B., Sakhardande, A., & Blakemore, S.-J. (2018). *The Matrix Reasoning Item Bank (MaRs-IB)*. https://osf.io/g96f4/

De Vos, T. (1992). *Tempo-Test-Rekenen: Test voor het vaststellen van het rekenvaardigheidsniveau der elementaire bewerkingen (automatisering) voor het basisen voortgezet onderwijs: Handleiding*. Berkhout.“Swets & Zeitlinger.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11.

Khanolainen, D., Van Bergen, E., Koponen, T., Salminen, J., Tolvanen, A., & Torppa, M. (2023). *The Math Ability Rating Questionnaire (MARQ) for Adults: Validity and Reliability*. OSF Registries. https://doi.org/10.17605/OSF.IO/Q2TF8

Khanolainen, D., Van Bergen, E., Tolvanen, A., Koponen, T., Salminen, J., & Torppa, M. (2025). *Math Ability Rating Questionnaire (MARQ) for Adults: Validating a New Self-report Measure*. PsyArXiv. https://doi.org/10.31234/osf.io/nsdqh_v1

Kievit, R. A. (2023, December 19). *Open Science project for Free Cognitive Testing Stimuli | Radboud University*.

R Core Team. (2024). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Schlichting, L. (2005). *Peabody Picture Vocabulary Test-III-NL. Nederlandse versie. Handleiding.* Amsterdam: Harcourt Test Publishers.

Appendix

TTR – version 1 (online home)

	1+1 =____		13+ 4 =____		17+81 =____
	2+1 =____		7 +12 =____		48+37 =____
	3+0 =____		16+ 8 =____		21+68 =____
	4+1 =____		4 +15 =____		67+24 =____
5	2+3 =____	*25*	17+ 3 =____	*45*	77+19 =____
	7+2 =____		6 +15 =____		55+38 =____
	3+5 =____		18+ 5 =____		42+29 =____
	0+7 =____		3 +14 =____		15+78 =____
	2+5 =____		17+ 8 =____		23+19 =____
10	4+6 =____	*30*	7 +16 =____	*50*	63+28 =____
	6+3 =____		17+16 =____		39+35 =____
	4+3 =____		22+13 =____		72+17 =____
	8+2 =____		19+32 =____		53+38 =____
	3+6 =____		34+15 =____		26+45 =____
15	5+2 =____	*35*	28+27 =____	*55*	66+25 =____
	3+8 =____		23+38 =____		18+77 =____
	5+7 =____		39+46 =____		56+35 =____
	2+6 =____		65+33 =____		41+58 =____
	7+5 =____		76+18 =____		33+39 =____
20	9+4 =____	*40*	54+27 =____	*60*	47+46 =____

Aantal gemaakte plussommen: ……

Aantal fout: ……

_____ –

Ruwe score plussommen: ……

	2-1 =____		18- 6 =____		69-14 =____
	3-2 =____		15- 3 =____		91-22 =____
	4-2 =____		16- 8 =____		43-16 =____
	3-0 =____		13- 2 =____		75-62 =____
5	5-2 =____	*25*	19- 7 =____	*45*	88-49 =____
	8-3 =____		28- 5 =____		53-29 =____
	6-0 =____		21- 9 =____		86-57 =____
	9-2 =____		27- 7 =____		65-41 =____
	7-5 =____		25- 8 =____		93-25 =____
10	8-6 =____	*30*	26- 9 =____	*50*	82-54 =____
	7-4 =____		35-17 =____		71-42 =____
	8-7 =____		48-23 =____		48-29 =____
	7-5 =____		26-19 =____		87-79 =____
	8-3 =____		44-32 =____		99-82 =____
15	6-5 =____	*35*	23-18 =____	*55*	52-37 =____
	15-3 =____		73-48 =____		35-28 =____
	13-7 =____		54-37 =____		46-27 =____
	18-8 =____		87-43 =____		83-65 =____
	16-9 =____		67-49 =____		71-43 =____
20	17-4 =____	*40*	43-27 =____	*60*	86-68 =____

Aantal gemaakte minsommen: ……

Aantal fout: ……

_____ –

Ruwe score minsommen: ……

TTR - version 2 (online lab)

	1+2 =____		14+ 3 =____		15+83 =____
	0+2 =____		7 +11 =____		45+39 =____
	1+1 =____		12+ 9 =____		22+63 =____
	3+2 =____		10+ 8 =____		66+28 =____
5	2+2 =____	*25*	15+ 5 =____	*45*	79+13 =____
	6+3 =____		8 +13 =____		34+49 =____
	4+5 =____		16+ 7 =____		26+39 =____
	0+8 =____		4 +12 =____		77+18 =____
	2+6 =____		18+ 5 =____		32+29 =____
10	3+7 =____	*30*	9 +17 =____	*50*	54+18 =____
	7+2 =____		15+16 =____		49+47 =____
	6+3 =____		21+15 =____		16+63 =____
	5+5 =____		17+35 =____		44+29 =____
	3+4 =____		31+14 =____		35+24 =____
15	2+6 =____	*35*	23+29 =____	*55*	77+24 =____
	3+ 9 =____		21+37 =____		17+66 =____
	4+10 =____		35+49 =____		45+37 =____
	2+ 7 =____		61+37 =____		52+39 =____
	5+ 9 =____		79+12 =____		44+48 =____
20	9+ 3 =____	*40*	59+26 =____	*60*	36+35 =____

Aantal gemaakte plussommen: ……

Aantal fout: ……

_____ –

Ruwe score plussommen: ……

	3-1 =____		15- 8 =____		68-13 =____
	4-2 =____		14- 6 =____		95-24 =____
	5-3 =____		16- 9 =____		44-26 =____
	2-0 =____		15- 4 =____		64-28 =____
5	2-2 =____	*25*	18- 6 =____	*45*	99-32 =____
	9-5 =____		25- 4 =____		64-19 =____
	8-0 =____		24- 7 =____		96-47 =____
	7-3 =____		20- 3 =____		55-31 =____
	9-1 =____		22- 9 =____		84-45 =____
10	8-7 =____	*30*	23- 5 =____	*50*	72-44 =____
	8-5 =____		29-11 =____		70-41 =____
	5-4 =____		25-17 =____		48-22 =____
	7-1 =____		32-18 =____		77-69 =____
	8-5 =____		34-17 =____		99-84 =____
15	9-3 =____	*35*	28-12 =____	*55*	63-34 =____
	11-5 =____		75-36 =____		33-17 =____
	13-6 =____		55-34 =____		48-26 =____
	12-7 =____		82-43 =____		89-55 =____
	16-6 =____		62-39 =____		81-55 =____
20	18-3 =____	*40*	47-24 =____	*60*	96-77 =____

Aantal gemaakte minsommen: ……

Aantal fout: ……

_____ –

Ruwe score minsommen: ……

TTR – version 3 (pen/paper lab)

	2+1 =____		12+ 5 =____		56+16 =____
	4+0 =____		8+11 =____		47+25 =____
	3+1 =____		14+ 7 =____		19+72 =____
	4+1 =____		9+11=____		63+29 =____
5	2+3 =____	*25*	14+ 4 =____	*45*	56+23 =____
	7+0 =____		8 +14 =____		41+23 =____
	4+2 =____		6 +17 =____		22+71 =____
	2+5 =____		9 +13 =____		60+34 =____
	3+4 =____		18+ 6 =____		42+56 =____
10	6+2 =____	*30*	5+16 =____	*50*	51+21 =____
	8+1 =____		17+15 =____		66+28 =____
	3+5 =____		23+12 =____		82+17 =____
	7+3 =____		22+28 =____		38+47 =____
	5+4 =____		35+13 =____		71+22 =____
15	10+1 =____	*35*	27+26 =____	*55*	36+58 =____
	9+ 2 =____		25+33 =____		58+33 =____
	3+11 =____		43+39 =____		41+45 =____
	6+ 6 =____		63+35 =____		62+34 =____
	5+ 3 =____		78+16 =____		38+36 =____
20	8+ 4 =____	*40*	53+28 =____	*60*	76+17 =____

Aantal gemaakte plussommen: ……

Aantal fout: ……

_____ –

Ruwe score plussommen: ……

	3-1 =____		20- 7 =____		68-15 =____
	5-3 =____		14- 3 =____		92-21 =____
	6-3 =____		18- 9 =____		44-17 =____
	4-1 =____		12-3 =____		77-63 =____
5	4-0 =____	*25*	15- 8 =____	*45*	89-52 =____
	9-4 =____		27- 3 =____		54-28 =____
	8-1 =____		22- 8 =____		85-56 =____
	9-3 =____		30- 8 =____		63-41 =____
	8-6 =____		28- 7 =____		84-53 =____
10	7-5 =____	*30*	26- 8 =____	*50*	91-27 =____
	9-6 =____		36-16 =____		72-41 =____
	8-7 =____		47-24 =____		49-21 =____
	7-4 =____		38-19 =____		88-59 =____
	9-8 =____		43-31 =____		98-83 =____
15	6-2 =____	*35*	24-17 =____	*55*	55-38 =____
	14-5 =____		74-47 =____		37-26 =____
	15-6 =____		56-39 =____		48-23 =____
	19-7 =____		85-42 =____		81-63 =____
	18-9 =____		70-48 =____		73-42 =____
20	19-5 =____	*40*	42-26 =____	*60*	85-67 =____

Aantal gemaakte minsommen: ……

Aantal fout: ……

_____ –

Ruwe score minsommen: ……

MARQ – Finnish, English and Dutch translations

Item	English	Finnish	Dutch	(Dutch) Scale
1	How good do you think your mathematical skills are compared to people your age?	Arvioi omia taitojasi ikäisiisi ihmisiin verraten. Kuinka hyvät taidot sinulla mielestäsi on?: Matematiikassa	Hoe zou u uw eigen rekenvaardigheid beoordelen in vergelijking met leeftijdsgenoten?	1–5 (Onder gemiddeld … Boven gemiddeld)
2	Think back to your own school days and evaluate your mathematical skills in relation to other students in your class.	Muistele vielä omaa kouluaikaasi ja arvioi seuraavia taitojasi suhteessa luokkasi muihin oppilaisiin. Kuinka hyvä olit matematiikassa?	Denk nog eens terug aan uw eigen schooltijd en beoordeel uw vaardigheden in vergelijking met uw klasgenoten. Hoe goed was u in rekenen?	1–5 (Slecht … Erg goed)
3	Did you experience any difficulties in learning math in elementary school?	Oliko sinulla vaikeuksia oppia matematiikkaa alakoulussa?	Had u moeite met leren rekenen op de basisschool?	1–5 (Helemaal geen moeite … Heel veel moeite)
4	How was your math performance compared to your classmates in elementary school?	Miten vertaisit omia matematiikan taitojasi alakoulun luokkatoveriesi taitoihin?	Hoe was uw rekenvaardigheid vergeleken met die van uw klasgenoten op de basisschool?	1–5 (Boven gemiddeld … Onder gemiddeld)
5	How do you rate your math skills now compared to people your age with a comparable education level?	Kuinka vertaisit nykyisiä matematiikan taitojasi toisten samanikäisten ja saman koulutuksen omaavien taitoihin?	Hoe zou u uw huidige rekenvaardigheid vergelijken met die van leeftijdsgenoten met een vergelijkbare opleiding?	1–5 (Boven gemiddeld … Onder gemiddeld)
6	Did you experience any difficulties learning the multiplication table in elementary school?	Oliko sinun vaikeaa oppia kertotaulu alakoulussa?	Vond u het moeilijk om de tafels te leren op de basisschool?	1–5 (Helemaal niet moeilijk … Heel moeilijk)
8	Looking back at your childhood… Did you have trouble learning how to do mental addition and subtraction?	Kuinka sinä pärjäsit matematiikassa lapsena? … Oliko sinulla vaikeuksia oppia laskemaan yhteen- ja vähennyslaskuja	Had u moeite met het leren optellen en aftrekken in uw hoofd?	1–5 (Helemaal niet … Heel veel)

		mielessä?		
9	Did you have trouble learning how to do multiplications?	Oliko sinulla vaikeuksia oppia kertolaskuja?	Had u moeite met het leren van vermenigvuldigen?	1–5 (Helemaal niet … Heel veel)
11	Did you have trouble learning to count money?	Oliko sinulla vaikeuksia oppia laskemaan rahoilla?	Had u moeite met het leren rekenen met geld?	1–5 (Helemaal niet … Heel veel)
12	Did you have difficulties learning and measuring units of measure (e.g. mm, cm, dm...)?	Oliko sinulla vaikeuksia mittayksiköiden oppimisessa ja mittaamisessa (esim. mm, cm, dm…)?	Had u moeite met het leren van meten en maateenheden (bijv. mm, cm, dm…)?	1–5 (Helemaal niet … Heel veel)
13	How much extra support did you need to learn to count in the first years of school (addition, subtraction and multiplication)?	Kuinka paljon ylimääräistä tukea tarvitsit oppiaksesi laskemaan ensimmäisinä kouluvuosina (yhteen-, vähennys- ja kertolasku)?	Hoeveel extra hulp had u nodig bij het leren rekenen in de eerste schooljaren (optellen, aftrekken en vermenigvuldigen)?	1–5 (Helemaal niet … Heel veel)
15	…mentally calculate the exact total of the purchases.	laskea mielessä ostosten tarkan loppusumman.	De totale prijs van aankopen in mijn hoofd uit te rekenen.	1–5 (Helemaal niet moeilijk … Heel erg moeilijk)
16	…calculate the price of a discounted product (e.g. calculating the final price of a product when it is 40% off).	laskea alennetun tuotteen hinnan (esim. tuotteen lopullisen hinnan laskeminen, kun se on 40% alennuksessa).	De prijs van een afgeprijsd product te berekenen (bijv. de uiteindelijke prijs uit te rekenen bij 40% korting).	1–5 (Helemaal niet moeilijk … Heel erg moeilijk)
17	…calculate the final price of the product if you pay in instalments.	laskea tuotteen lopullisen hinnan, jos maksat sen erissä.	De uiteindelijke prijs van een product te berekenen wanneer je het in termijnen betaalt.	1–5 (Helemaal niet moeilijk … Heel erg moeilijk)
18	…calculate the amount of salary left after deducting taxes and side costs.	laskea käteen jäävän palkan suuruuden verojen ja sivukulujen vähentämisen jälkeen.	De hoogte van het overgebleven salaris te berekenen na het aftrekken van belastingen en extra kosten.	1–5 (Helemaal niet moeilijk … Heel erg moeilijk)
23	…converting length measurements in	pituusmittojen muuntaminen mielessä	Lengtematen in mijn hoofd om te rekenen	1–5 (Helemaal niet moeilijk …

	mind (e.g. centimetres to metres or millilitres to centimetres, etc.)	(esim. sentit metreiksi tai millit senteiksi).	(bijv. centimeters naar meters of millimeters naar centimeters).	Heel erg moeilijk)
26	…calculate the required amount of fuel for a certain distance if you know the average consumption rate of the car (l/100 km)	laskea tarvittava polttoainemäärä tietylle matkalle, jos tiedät auton keskimääräisen kulutuksen (l/100 km).	De benodigde hoeveelheid brandstof te berekenen voor een bepaalde afstand als je het gemiddelde verbruik van de auto kent (l/100 km).	1–5 (Helemaal niet moeilijk … Heel erg moeilijk)
29	I don't have a math brain	Minulla ei ole ”matikkapäätä”.	Ik heb geen “wiskundeknobbel”.	1–5 (Helemaal niet waar … Helemaal waar)
30	I'm good at math	Olen hyvä matematiikassa.	Ik ben goed in rekenen.	1–5 (Helemaal niet waar … Helemaal waar)