Psychological Test and Assessment Modeling, Volume 60, 2018 (4), 451-492

Scrambled Adaptive Matrices (SAM) – a new test of eductive ability

Balázs Klein[1], John Raven[2] & Szilvia Fodor[3]

Abstract

Inspired by the Raven's Progressive Matrices a new, innovative, IRT-based online adaptive test, the Scrambled Adaptive Matrices (SAM) has been developed and used in talent identification projects both in educational and work settings with more than 15 000 participants. The current article introduces the test and shows results on reliability and validity as well as response time, motivational concerns, adaptivity, security and the effects of external variables such as socio-economic status, age, gender and pre-selection. The data show that the newly developed instrument is a feasible, reliable and valid tool for ability assessment and talent identification in projects of all sizes.

Keywords: computerized adaptive testing, intelligence, item response theory, Raven Progressive Matrices, validity

[1] *Correspondence concerning this article should be addressed to:* Balázs Klein, University of Pécs, Research Center for Labor and Health Sciences, Hungary; email: balazs.klein@gmail.com

[2] United Kingdom

[3] University of Debrecen, Hungary

The purpose of this study was to investigate and report the properties of a new, IRT-based online adaptive test of eductive ability – the Scrambled Adaptive Matrices (SAM). Inspired by the Raven's Progressive Matrices the test was developed by Klein in 2015 and has been used in talent identification projects both in educational and work settings with more than 15 000 participants since. The current study introduces the test and shows results on reliability and validity as well as response time, motivational concerns, adaptivity, security and the effects of external variables such as socio-economic status, age, gender and pre-selection.

Theoretical background

In the beginning of the 20th century Spearman discovered that different cognitive abilities correlate highly with each-other and came to the conclusion that a common factor – the general cognitive ability ("g") - must be responsible for this phenomena (Spearman, 1924). He proposed that g consists of two very distinct abilities that complement each-other:

- "meaning making ability" which he called eductive ability, referring to the Latin "educere" word meaning draw or take out,
- and the ability to reproduce previously learned information, which he termed "reproductive ability" (Spearman, 1927).

Similarly to Spearman Cattel (1987) and Horn (1994) also proposed a two factor model of intelligence:

- the fluid intelligence is the ability to solve new problems (independently from knowledge accumulated in the past) while
- crystallized intelligence is the ability to recall and utilize past knowledge and experience.

Though it is important to emphasize that reproductive ability is not a crystallized form of eductive ability and the two abilities have different origins, are affected by different aspects of the environment, are related to different brain functions predict different outcomes in life and change differently with age, in the literature "fluid intelligence" is often used as a synonym of "eductive ability" while "crystallized intelligence" is used for "reproductive ability" (Raven, 2008 a).

Measurement of eductive ability

While reproductive ability is usually measured by vocabulary scales, eductive ability is usually measured by some sort of diagrammatical test like the Raven Progressive Matrices (RPM) developed by J.C.Raven in 1936 (Raven, 1936).

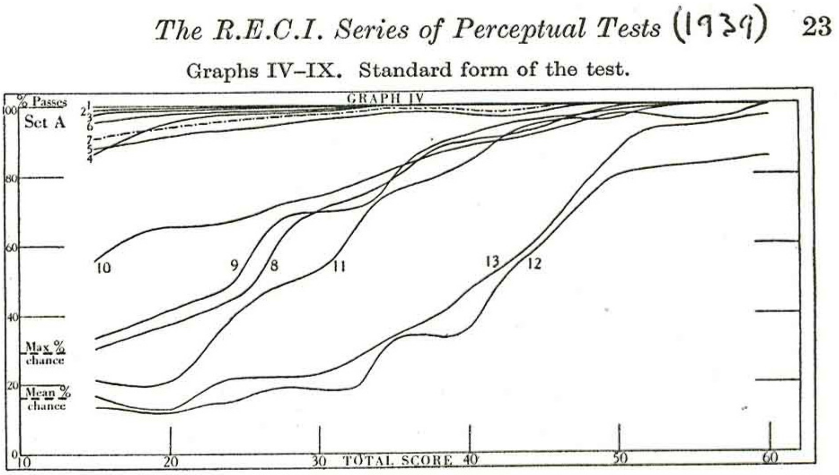
Raven's goal was to develop a test that

- requires no special training to administer,

- that can could be administered individually or in groups in a wide variety of settings (including homes, schools, and workplaces where there might be noise, potential “helpers”, and no quiet space),
- was easy to score and interpret,
- was theoretically sound,
- did not depend on the ability to read (e.g. written instruction) or write or even speak the same language as the administrator, and
- yielded a reliable score for people of all abilities at all ages from 4 to 80 years.

He developed the test primary for the purpose of the research he was engaged in at the time (Raven, 2000) although it quickly found application especially in the armed forces in the second world war.

At the time many people wanted to define "intelligence" as "the inherited part of mental ability". Raven’s goal was instead to create a test which made it possible to meaningfully investigate the issue of inheritance instead of building it into the definition. This was one of the good reasons for avoiding all use of the word "intelligence". Likewise, to use a more recent term, the existing measures at the time were arbitrary in the sense that scores on tests actually measuring a number of different things were arbitrarily combined together with essentially equal weights to yield a total score. By using item level measures – essentially item characteristic curves calculated and drawn by hand as shown in Figure 1 – he demonstrated that whatever was being measured by the more difficult items was "the same thing" as that measured by the easier ones - and in that sense the composition of the scale was non-arbitrary. At the time it was not sufficiently appreciated that being non arbitrary also implied having an interval scale so that the differences between the



The R.E.C.I. Series of Perceptual Tests (1939) 23

Graphs IV–IX. Standard form of the test.

Figure 1:
Hand-drawn item characteristic curves of the SPM

raw scores at different points in the scale meant the same thing. Failure to take steps to achieve this has resulted in endless unjustified claims about the development of eductive ability and effectiveness of "remedial" programs among the more and less able. These defects have been rectified in the *Standard Progressive Matrices **Plus*** (SPM+).

Today the RPM is one of the best known instrument to measure eductive ability and possibly the most pervasive non-individual test today (Kaplan – Saccuzzo, 2012). Its central role among cognitive tests are shown in Figure 2.

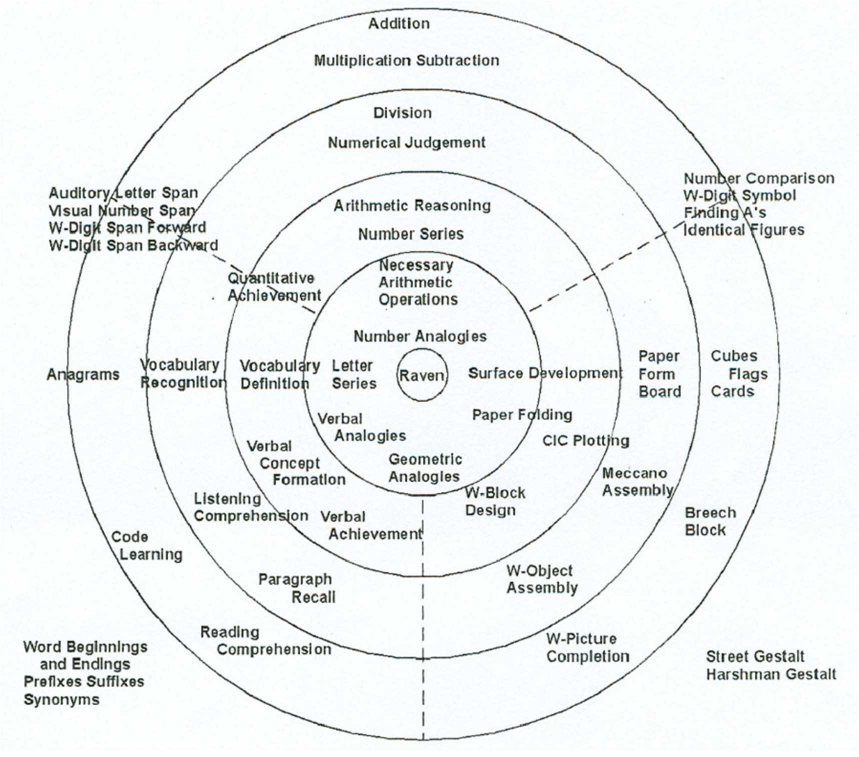


Figure 2:
The central role of the RPM drawn from the intercorrelations of different cognitive ability tests (Snow, Kyllonen & Marshalek, 1984)

Description of the RPM

In the RPM the task of the participant is to find the missing piece of a 3 by 3 matrix among the given options so that the pieces – containing black-and-white geometric figures – follow each other both horizontally and vertically in a logical order. Figure 3 shows an item from the Advanced Progressive Matrices.

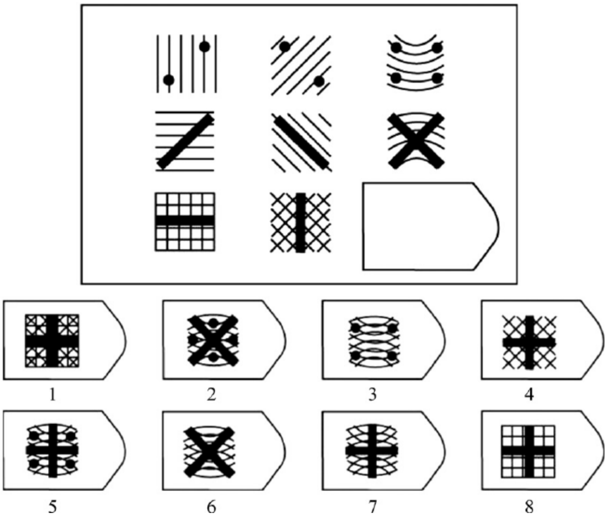
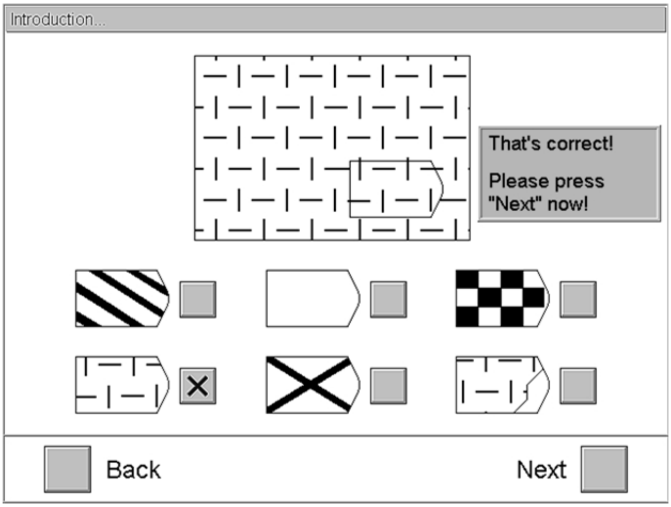


Figure 3:
An example item from the Advanced Progressive Matrices test (Pearson, 2018)

In the past years online tests – similar or identical to the RPM – have appeared on the internet in large numbers. Most of these tests however concentrate only in advancing the delivery mechanism of the original test by making test taking and scoring easier. Figure 4 shows an example of the online version of the RPM.


Introduction...

That's correct!

Please press "Next" now!

Figure 4:
Sample item of the online version of the RPM (Schuhfried, 2016)

The way forward

The RPM aims and claims to measure eductive ability (Raven, 2008 b) as if this were a general trait across all areas of thinking, while it is probably not. As Spearman noted "the question is not 'how well can they think' but 'what do they tend to think about'". Whether "thinking" – let alone eductive ability – as displayed while crafting a painting is the same thing as attending to Matrix problems remains to be seen. While the ability to create disruption in a classroom seems unlikely to rely on the same processes as the ability to find a solution to mathematical problems our competence studies seem to indicate this: effective disruptors of school classes study the effects of their actions and learn from them, try experimental variations, turn their emotions into the task etc. But would all of that show up if they were confronted with the RPM? So what is eductive ability as an abstracted notion? And does the attempt to focus attention on it deflect attention from other things? Should we be concerned by the measurement of intelligence or are there more important factors that could contribute to the survival of mankind on Earth?

In any case if we are to measure eductive ability we should certainly do it well. Many look-alikes of the *Progressive Matrices* have been published (Web, 2018), but a large proportion of them were developed without a clear understanding of the theoretical basis of the tests or the scaling procedures used in their development. Very few have involved the creation of theoretically-based variants of the items. Many critiques have stemmed from the attempt to fit Classical test theory to sets of item statistics. Many have been based on attempts to factor-analyse the correlations between the items. As anyone familiar with "the Rasch model" knows, and is more specifically demonstrated in Raven & Fugard (2008) this results in claims that the tests measure a number of different things supposedly identified as different "factors". Yet what emerges from a plot of the Item Characteristics Curves resulting from application of an appropriate Item Response Theory program (Figure 5) is clear evidence that the abilities required to solve the more difficult items are built upon, incorporate, and extend the abilities needed to solve the easier ones (Raven, Prieler & Benesch, 2008).

They do not involve suddenly emerging "new abilities". There are no "metamorphoses" or "leaps in development"[4]. What is more, despite often huge differences in the average scores of different cultural groups the test works, and works in the same way, among people from different socio-economic and ethnic backgrounds (Raven, 2008). The extraordinary graph on Figure 6 showing the test characteristic curves derived from large samples of blacks and whites in South Africa (Taylor, 2008) makes this abundantly clear.

[4] For the sake of academic integrity, it is necessary to say that Figure 5 is somewhat misleading in that the parameters applied in the process of smoothing the curves have resulted in creating a somewhat misleading impression. A more realistic figure based on 3-Parameter Model is available in Raven, Prieler & Benesch (2008).

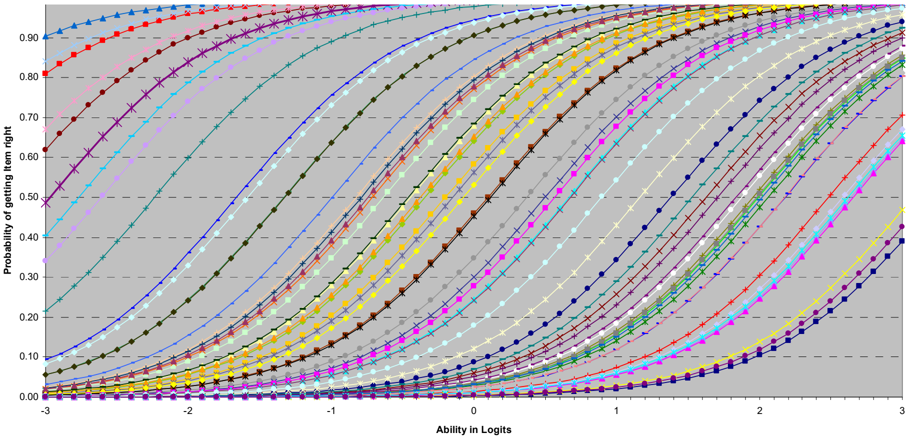


Figure 5:
Item characteristic curves of the SPM+ (Raven, Prieler & Benesch, 2008)

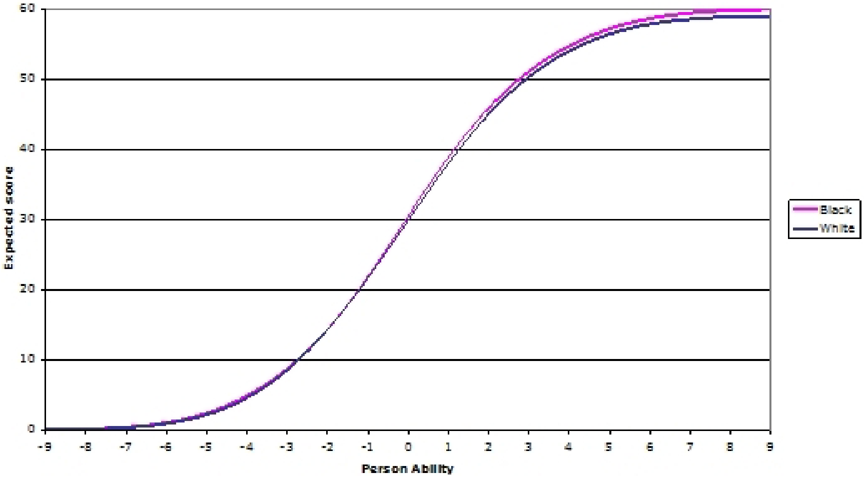


Figure 6:
Test characteristic curves for blacks and whites on the SPM (Taylor, 2008)

The Scrambled Adaptive Matrices (SAM)

The Scrambled Adaptive Matrices was developed by Klein to fulfill the need for a high quality online assessment test of eductive ability. The test can be used in ability assessment, talent identification or research projects with large number of participants on desktops or mobile instruments. Test administration can be done through the admin interface of the system or through computer application interface (API) by external applications.

The need to develop a new measurement of eductive ability

The Raven Progressive Matrices has been an excellent measurement tool of eductive ability in the past almost hundred years, yielding useful information in both practical and academic projects. The RPM however has a number of inherent properties that now with the advances of technology provide some opportunities for further development.

- The RPM is a paper and pencil test.
  Paper and pencil tests make running large, multi-thousand participate projects logistically challenging and expensive. This is the delivery issue that the online versions of the RPM mostly try to address.
- The RPM is a fixed test.
  The RPM – like all paper and pencil tests – is fixed, meaning that all test takers receive the same tasks. In today's world where information is easy to find the scoring key of the RPM (similarly to other fixed tests) can easily be found on the internet, thus making high-stake decisions based on its result risky. Though for some versions of the RPM a parallel version is available this offers only a partial solution to the problem.
- The RPM is long.
  The length of the RPM is between 36 and 60 items. Testing time varies, different versions have different time limits. There are frequent deviations in practice from the time limit set by the test manual. According to the instructions there is no time limit for the SPM or SPM+ and there is a 40 minutes time limit for the APM. In practice the SPM is often administered with a time limit. Though there is always a risk in introducing time limit to a power test a shorter test with more consistently enforced time limit has some advantages over the current situation.
- Secondary variables are not easily collected
  With computer based testing the we can easily and consistently collect secondary variables such as response time. This can possibly provide us important insights in academic studies. Currently this information is not collected, or only collected on the test level in most projects with the RPM.
- The RPM needs different versions for different ability levels.
  The Raven's tests have 6 versions in 4 difficulty level. These are shown in Table 1.

Table 1:
Versions of the Raven's test (Raven et al., 2000)

	Level			
Version	1	2	3	4
1	Coloured Progressive Matrices (CPM)	Standard Progressive Matrices (SPM)	Standard Progressive Matrices Plus (SPM+)	Advanced Progressive Matrices (APM)
2	Coloured Progressive Matrices Parallel (CPM parallel)	Standard Progressive Matrices Parallel (SPM parallel)		

For fixed tests to remain short different versions of the test had to be developed to optimally measure at different ability levels. Choosing a wrong (too easy or too difficult) test can largely increase the measurement error and cause a floor or ceiling effect. Therefore it is essential to choose the right test prior to testing – which always introduce an element of risk.

- The RPM uses only black and white images.
  Raven items – except in the CPM – are black-and-white while color can be more engaging to all ages and abilities.
- The RPM is timed for the whole test.
  When time limit is applied to the RPM it is applied to the whole test. Results when the test taker is running out of time may be affected by the lack of time. This is against the item-independence presumption of IRT and can distort item parameter estimations.
- The RPM is difficult to update in the long term.
  The RPM has been around in an unchanged format for almost a century now and so it provided invaluable insight into the changes of IQ over time. However looking into the future the further development of the paper and pencil based RPM is uncertain (e.g. though it has outstanding psychometric properties it took considerable time and effort to develop the SPM+ properties).
- The arrangement of the RPM is not optimal for handheld devices.
  With the task format of the RPM the options take up a lot of valuable screen space which can be a problem on handheld devices. The BiNona format of SAM addresses this issue.
- Reduced differentiation at the very high end of the ability range
  It has proved to be impossible to develop properly functioning items at the high end of the ability range in the RPM therefore differentiation among most able test takers is less than optimal.

– The “interface” is static – trial and error is only mentally possible.
  In the RPM the interface is static. The test taker simply marks the answer that she believes is correct and must mentally check the fit between the chosen option and the remaining of the matrix.

Task format

For obvious copyright reasons all tasks of the SAM test are new and different from the RPM tasks. Furthermore while preserving the fundamental problem situation of the RPM – which is identifying rules in chaos – , and the basic arrangement of the problem pieces – 3 by 3 matrices – the SAM tasks are not only graphically but also structurally different from the RPM tasks.

$\binom{9}{2}$ The task situation of the SAM test is to choose the two pieces of the 3 by 3 matrix that needs to be swapped for all pieces to be logically arranged both horizontally and vertically. Later this format was found to be useful in other tests as well. The name given to this format is BiNona from the latin bin (two) and nona (nine) words describing the fact that the task is to choose two pieces out of nine. The name of the test – Scrambled Adaptive Matrices – also refer to the new task situation where the pieces are originally scrambled and the test taker must order them correctly by swapping two pieces.

In the SAM test after choosing the two pieces the computer automatically swaps them so the result of the swap can be confirmed or rejected by the test taker by a Next or an Undo button. Typically 25 tasks are given with a time limit of 2 minutes. Completing all the tasks usually takes around 20 minutes to solve. When the time elapses the system automatically forwards the test taker to the next task – accepting any correct solution if already made. Candidates are therefore encouraged to take as much of their allocated time as needed instead of hurrying and giving an answer even if they are uncertain of their solution.

The following figure shows the three practice items of the test (Figure 7). For demonstration purposes the two tiles that needs to be swapped are marked with a double-headed arrow. The practice items show that both rule types and item difficulties can be widely different in the test. In the solutions it is visible that the background of the selected pieces is grey and their outlines are green. Since the swap is already made and (unless the swap is undone) further interaction is not possible the task becomes locked which is signaled by fading the task and placing locks in its four corners.

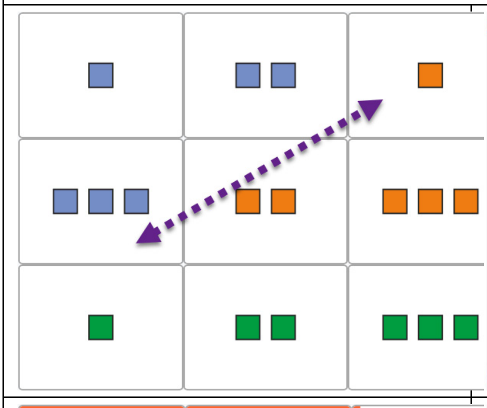
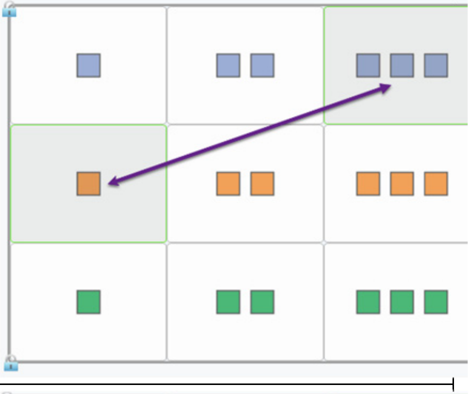
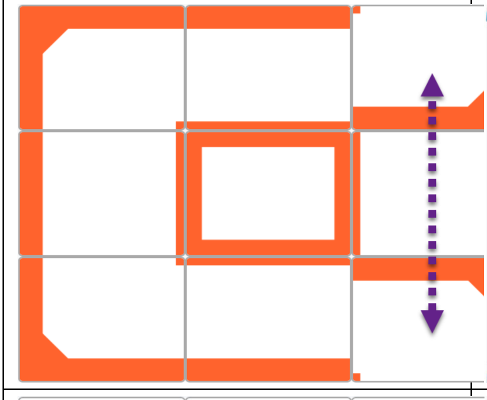
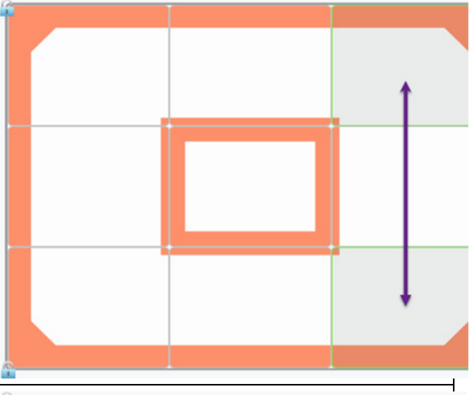
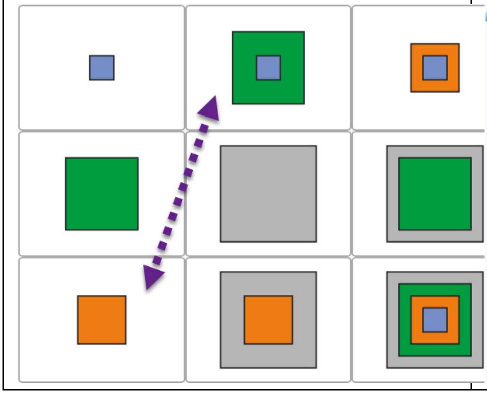
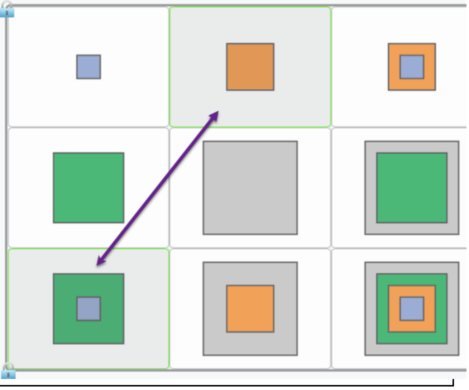
Task	Solution
Swap two tiles to make all tiles follow each other in a logical order both horizontally and vertically!	
	
	
	

Figure 7:
Practice items of the SAM test

This arrangement is a significant change compared to the RPM. The test taker encounters wrong or misleading information and rule making becomes an iterative process. In Neisser (1976) paradigm – shown in Figure 8 – the process looks like the following:

1. The first step is the exploration of the information provided.
2. The next step is the creation of a hypothetical rule which directs further attention.
3. The next step is using the collected information to confirm or reject the hypothesis.
4. In case of a rejected hypothesis a new one is formed and the process continues.

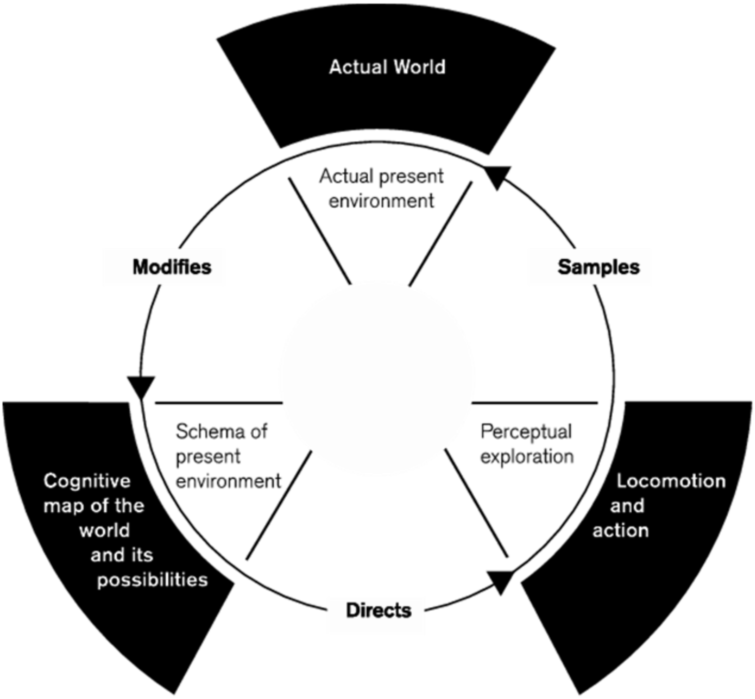

Actual World
Actual present environment
Modifies
Samples
Schema of present environment
Perceptual exploration
Cognitive map of the world and its possibilities
Locomotion and action
Directs

Figure 8:
The interaction of the environment, its exploration and its mental schemas (Neisser, 1976)

This new tasks setting significantly changes the type of tasks that work well in the test. While previously increasing the number of rules that effected the matrix inevitably increased the complexity and difficulty of the task in the new setting this is not the case. Here it is enough to identify which two pieces needs to be swapped by ANY of the rules operating on them.

Test taking procedure

The participant goes through the following procedure during test taking:

1. Login
   In case of supervised group testing login is usually done using preprinted codes handed out by the supervisor. In case of unsupervised testing login is usually done using a link sent out in an e-mail. The system provides facilities for both options.
2. Wait page for group testing (optional)
   In group testing projects it can be important that all test takers start at the same time. This page requires a code – provided by the test supervisor – to continue.
3. Welcome page
   The test starts with a welcome page. This is also the page from which test taking can be continued from a previously interrupted occasion.
4. Terms and Conditions page (optional)
   This page appears in some projects where test takers must accept the terms and conditions of usage.
5. Instruction page
   A page appears with detailed instructions that enable users to complete the test without supervision.
6. Practice item page
   A practice item that works the same way than the actual items, but is not scored.
7. Feedback page for the practice item
   The computer feeds back whether the answer to the practice item was correct or not. Some hints and insights are also provided. If the answer was not correct the user can look at the correct solution, but in any case must try the practice item again until he gives a correct response. The user can only proceed after giving correct answers to all 3 practice items.
8. Final instructions page
   Before starting the real test the user is presented with the final instructions. In group testing projects (optionally) the user must wait until the test supervisor gives the start code.
9. Item pages
   Usually 25 items displayed adaptively. Although discouraged test taking can be interrupted and continued. There is a 2 minutes time limit for each item.
10. Thanks page
    After the item pages the participation of the user is thanked and the test is closed.
11. Results page (optional)
    In some projects the result of the test taker is immediately fed back on a web page, in a downloadable document or both.

The measurement paradigm of SAM

In the background the SAM is using an IRT based scoring algorithm. IRT (item-response theory) is a measurement paradigm describing how we think about the development or scoring of tests. The most important difference between IRT and classical test theory (CTT) is that while CTT focuses on the test as a whole – as its name suggests – IRT considers the properties (e.g. difficulty) of each individual item (van Alphen et al., 1994). The theory states that the probability of a correct answer is dependent on the ability of the test taker and the properties of the items. IRT has been around since the 1950s but started to spread only in the 1970s with the increasing availability of computers.

IRT assumes that the measured trait is unidimensional, that the responses are independent (e.g. there is no learning effect or test level time constraint) and that they can be modelled using an item-response function.

The model used by SAM is based on the 2 parameter model. This means item characteristic curves describing the behavior of an item in relation to the performance on the test (like those on Figure 4) are using a difficulty parameter (the inflection point on the X axis) and a difficulty parameter the (the slope of the curve at the inflection point).

Adaptive testing (CAT)

Item parameters calculated by IRT methodology not only help us to achieve more precise estimation of the ability level of the test takers but enable us to get comparable results independently from the fact that different test takers received different items.

Having comparable results independently from the items displayed enable us to select tasks that yield the most information about the ability of the test taker. We achieve this using an iterative algorithm.

1. We give an estimate of the ability of the test taker. At the beginning of the testing this could be the mean ability (of the test takers group).
2. From the item bank we select the task that will yield the most information – a task with high discrimination and close to the ability level of the test taker.
3. Based on the right or wrong answer that we receive we recalculate the ability estimate and the process can restart.

By using this method we can get significantly more information about the ability of the test taker than by using a fixed set of items therefore we can get more precise estimation with the same amount of tasks or reduce the length of the test while maintain its precision. Computer adaptive testing (CAT) can usually reduce the length of tests by half (Weiss - Kingsbury, 1984).

Figure 9 shows how the estimates to the ability changes during a test and how that effects the selection of the next task. In case of a correct answer the algorithm chooses a more difficult tasks, while in case of an incorrect answer an easier task. Meanwhile the estimate to the ability is becoming more and more precise. This technology helps reducing test length while ensuring that everyone receives a different test.

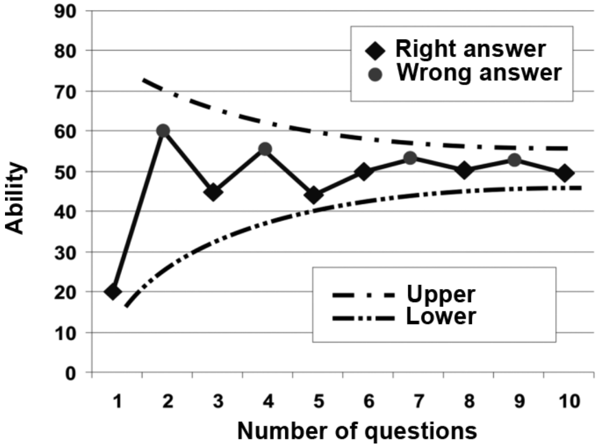


Figure 9:
Adaptive testing and the error of ability estimation

The usual problem with adaptive testing is that since the computer needs to calculate the next item based on the response given after each response the user must wait for the next task. Our algorithm in SAM could eliminate this waiting time between tasks creating a flawless user experience – which was found to be especially important with children.

Overall adaptive testing helps us to

- display different individualized tests for every test-taker while preserving comparability,
- achieve very high levels of reliability in an extremely wide range of ability,
- shorten our test,
- constantly enlarge and enhance our item bank.

Test development under the IRT paradigm

Test development under the IRT paradigm is very different from classical test development. While classical ability tests have a limited number of items that are usually prepared within a limited time period before the publication of the test with an IRT based adaptive test a large item bank is preferable. In case of a test used with a significant number of test takers of a wide ability range the item bank may consists hundreds of items with new items being constantly added and trialed. Table 2 describes the most important differences between classical and IRT based test development in general.

Table 2:
Differences between traditional and IRT test development

	Traditional	IRT
Item bank	max. 60	250+
Normgroup	>1k	>10k
Technology	paper	CAT
Updates	minimal	continuous

The item bank

Test adaptivity is based on a large and high-quality item-bank. The item bank of SAM currently consists of more than 300 active items and many more in the trial phase. The difficulty range of the items go beyond 3 standard deviation both below and above the average. To ensure quality only items above 0.4 discrimination are retained in the item bank.

The SAM test employs a 2 parameter IRT model for both scoring and item selection. Since the algorithm always tries to select the most informative items tasks with higher discriminatory power are displayed more often. The SAM test does not rely on artificial protocols against item overexposure, instead new items are constantly added to the item bank and trialed. This automatically reduces the frequency of individual item use.

Figure 10 shows the difficulty and discrimination of the items currently in the item bank. One can see from the image that there are many items on every difficulty level with

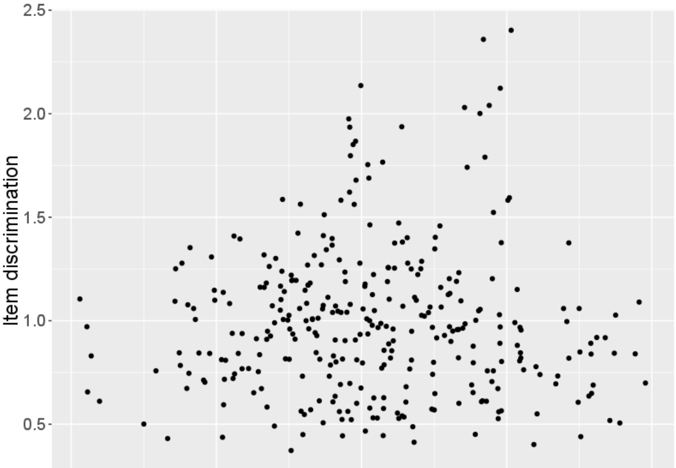


Figure 10:
The difficulty and discrimination of the items in SAM

sufficient discrimination power. It is also visible that in the task setting of SAM it was possible to create highly discriminative items even at the very high end of the ability spectrum. In fact some of the best performing items of the test are around 2 standard deviation above the average, which means that we have good tasks even for the top 2% of the population and above.

Results

In the following section we describe the projects in which the test was so far used and the different results demonstrating the features of the test.

List of projects

The SAM test is somewhat unique in the sense that it is extensively used both in the world of work and education.
Table **3** shows the major projects carried out until the end of 2018 September containing cc. 15 000 test takers. These test takers have very different backgrounds. There are students from elementary schools, secondary schools and universities as well as university graduates, professionals and adults with little prior education. The language of the online instructions were English or Hungarian. Results with unrealistically low testing times (less than 10 seconds per response) are already filtered out of the data. The columns and projects of the table are individually described below.

Column descriptions

The columns contain the following information:

- ProjektID – The unique reference name to the project.
- Language – The language of the instructions. Most projects were carried with Hungarian online instructions while some with English.
- Supervised – Was a supervisor present during test-taking? Some projects (especially school projects) were carried out on the premises of the institution with a supervisor present while others were registration based but completed unsupervised at home or completely open to anyone wishing to participate.
- Country – Typical country of residence of the test taker. The large majority of the projects were carried out in Hungary. In case of an open project the country given is the one from which most test takers came from. There is one project where the country is non-disclosed (ND) to preserve the anonymity of the institution. A Hungarian-speaking university outside of Hungary could be identifiable given the country.
- HighMotiv – Were test takers highly motivated to perform well? In some projects the outcome of the test is taken more seriously than in others. In some settings a high performance could contribute to a job offer or indicate eligibility to membership. In

these projects there was a good chance that participants were more motivated to perform well than in other cases.

- NumTests – The number of completed tests in the project. Normally the number of people completing a SAM test. The only notable exemption is the test-retest study where participants filled out the test multiple times so the number of completed test is 2-3 times higher than that of the test takers. This is marked by an asterix* in the table.
- Age – The average age of the test taker at test completion in years.
- Male% – The percentages of males in the group.
- RespSpeed – The average of response speed for an item in seconds. Response speeds are individually recorded for every response. At the completion of the test a median response speed is calculated for the whole occasion. The data in the table is the average of these medians in the project.
- Correct% – The average percent of correct answers in the group.
- Sem – The average of the individual standard errors of measurements.
- Theta – The average ability in the group.

Table 3:
Projects with the use of SAM

ProjektID	Language	Supervised	Country	HighMotiv	NumTests	Age	Male %	RespSpeed	Correct %	Sem	Theta
SamTrialHu_1	hu	n	hu	n	1410	30	13	39	46	0.47	0.22
SamTrialEn_1	en	n	ph	n	7450	24	38	37	44	0.43	-0.18
Consultancy_1	hu	n	hu	y	102		64	61	55	0.30	0.79
SamTrialHu_2	hu	n	hu	n	1607	38	34	47	54	0.31	0.43
IT_1	hu	n	hu	y	151		84	64	54	0.30	0.93
UniStaff_1	hu	n	hu	y	111		75	64	50	0.30	0.44
ElemKids_1	hu	y	hu	n	1420	11	50	35	50	0.30	-0.38
ElemKids_2	hu	y	hu	n	295	(11)		34	47	0.30	-0.24
UniStudents_1	hu	n	ND	n	472		28	50	52	0.30	0.31
Agency_1	hu	n	hu	y	254		60	62	56	0.30	0.52
Production_1	hu	y	hu	y	63	(35)	67	68	49	0.28	0.40
ElemKids_3	hu	n	hu	n	492	11	50	31	45	0.31	-0.67
TalentSite_1	hu	n	hu	n	114	26	42	50	53	0.29	0.45
3sam	hu	n	hu	n	76*			51	61	0.25	1.06
ElemKids_4	hu	y	hu	n	182	13	38	37	53	0.29	0.09
UniStudents_2	hu	n	hu	n	402			49	58	0.27	0.89
ElemKids_5	hu	y	hu	n	525	11	54	33	46	0.30	-0.43
Mensa_1	hu	n	hu	y	68	35	53	66	63	0.29	1.83
SecKids_1	hu	n	hu	n	152	15		41	55	0.27	0.53
Mensa_2	en	n	ch	y	102	24	67	64	58	0.28	1.30

Project descriptions

Below are the descriptions of each individual project that appear in the table above.

First Hungarian trial project (SamTrialHu_1)

This was an openly accessible and advertised project where anybody could fill out the instrument and received instant feedback in Hungarian. The feedback contained warning that the instrument was under development. The aim of this project was to establish the basic properties of the instrument. Since at this point we did not yet have item parameters the items were randomly assigned to test takers. This is the cause of the high SEM values. Probably because of the economics of the advertisement the 1410 test takers were mostly women (87%) with the average age of 30 years, graduated from secondary school, but not working at the time of completing the test.

First English trial project (SamTrialEn_1)

This was an openly accessible and advertised project, where anybody could fill out the instrument and received instant feedback in English. The feedback contained warning that the instrument was under development. The aim of this project was to establish the basic properties of the instrument. Since at this point we did not yet have item parameters the items were randomly assigned to test takers. This is the cause of the high SEM values. Test takers for this project came from different countries, but – probably because of the economics of the advertisements – the 7450 test takers mostly came from the Philippines, were mostly women (62%) with the average age of 24 years, graduated from college, but not working at the time of completing the test.

Hungarian consultancy firm (Consultancy_1 [[loxon]])

This project was part of the hiring process for a high-prestige Hungarian consultancy firm. Testing in this and all subsequent projects were done adaptively based on the IRT parameters calculated using previous data. The drop in the Standard Error of Measurement is striking compared to previous random item selection. Test takers in this project were mostly males (64%) soon after graduating from University. Since the job offer partly depended on the test result participants in this project are presumed to be highly motivated to perform well on the test.

Second Hungarian trial project (SamTrialHu_2)

This was the second Hungarian trial project. The project was openly accessible and advertised. Testing in this project was already done adaptively based on the IRT parameters calculated using previous data. Similarly to the other advertised open access projects test takers were mostly females (66%) graduated from secondary school and typically working at a white collar job.

Hungarian IT company (IT_1 [[nware]])

This project was part of the hiring process for a selective Hungarian IT firm looking for programmers. Test takers were overwhelmingly males (84%) soon after graduating from University. Since the job offer partly depended on the test result participants in this project are presumed to be highly motivated to perform well on the test.

Selection of Hungarian University staff members (UniStaff_1 [[elte]])

This project was part of the hiring process for a major Hungarian University looking for staff members at various administrative positions. Test takers were mostly males (75%). Since the job offer partly depended on the test result participants in this project are presumed to be highly motivated to perform well on the test.

Elementary school kids in a Hungarian region 1 (ElemKids_1)

In this project kids at the fifth grade of the elementary school were tested in several schools in a Hungarian region. Testing was done in groups within the schools under supervised conditions. The average age of the test takers were 11 years. Boys and girls participated in equal numbers in this project.

Elementary school kids in a Hungarian region 1 (ElemKids_2)

This project is considered to be the continuation of ElemKids_1.

University students in Central Europe (UniStudents_1)

In this project Hungarian-speaking student in a University of a neighboring country of Hungary were tested. Interestingly test takers were mostly (72%) women.

Assessment company (Agency_1)

This group of test takers don’t belong to a particular project but to a particular agency servicing different companies in their hiring decisions for positions ranging from basic catering jobs to senior management positions. Test takers in this group were mostly men (60%).

Line managers in a modern factory (Production_1)

In this project a modern Hungarian factory was looking to promote existing factory workers into line managerial position. Testing was a part of the selection process. Test takers in this group were mostly men (67%).

Elementary school kids in Hungary 1 (ElemKids_3)

This was a pilot project for a large representative testing of school kids at Grade 5 aimed at talent identification. The sample was aimed to be as closely representative to the total Hungarian school system as possible in terms of location, settlement type or previous competency scores. Testing was done supervised in the schools and included one complete class at grade 5 from every selected school. The average age in the group was 11 years. Girls and boys were equally represented in the sample.

Talent site (TalentSite_*1)*

This was an openly available but barely advertised project on a Hungarian web portal for high ability children.

Test-retest study (3sam)

In this project psychology students of a Hungarian University filled out the test preferably three times so that we could look at test-retest and learning effects. Therefore in this project the number of tests (shown in the table) is two-three times of that of the test takers.

Elementary school kids of 7th grade (ElemKids_4)

In this project elementary school kids filled out the test in supervised settings. Their average age was around 13 years.

University pilot project (UniStudents_2)

This is a pilot project of another possibly large-scale University testing project. A large number of background variable were collected in the project which is not yet available at the time of this compilation. Testing was done in unsupervised setting.

Elementary school kids in Hungary 2 (ElemKids_5)

This project is the pair and continuation of the project ElemKids_3 for another batches of schools.

Hungarian Mensa members (Mensa_1)

Interested members of the Hungarian Mensa could fill out the test voluntarily in an unsupervised setting. The average age of the group is around 35 years and there were roughly equal number of man and women in the group.

Secondary school kids (SecKids_1)

Secondary school kids in a Hungarian region with the average age of 15 years and roughly equal number of boys and girls.

Chinese Mensa pretest (Mensa_2)

Candidates aspiring membership into the Chinese Mensa could fill out the test checking their chances to get in before participating in the actual admission testing. Testing was done unsupervised with English instructions. Their average age was 24 years and most (67%) of them were males.

Response time

Response times are collected during testing and aggregated within each test occasion into the median response time in seconds of all responses (Figure 11). Further analyses of response time will refer to this data as the response time of an individual. As can be seen in the table of projects the median response times are usually far below (ranging from 33 sec to 68 sec) the maximum allowed 120 sec. This suggests that the SAM is a power (as opposed to speed) test where the performance of the test taker would not or only minimally increase if they were allowed to spend more time on the tasks.

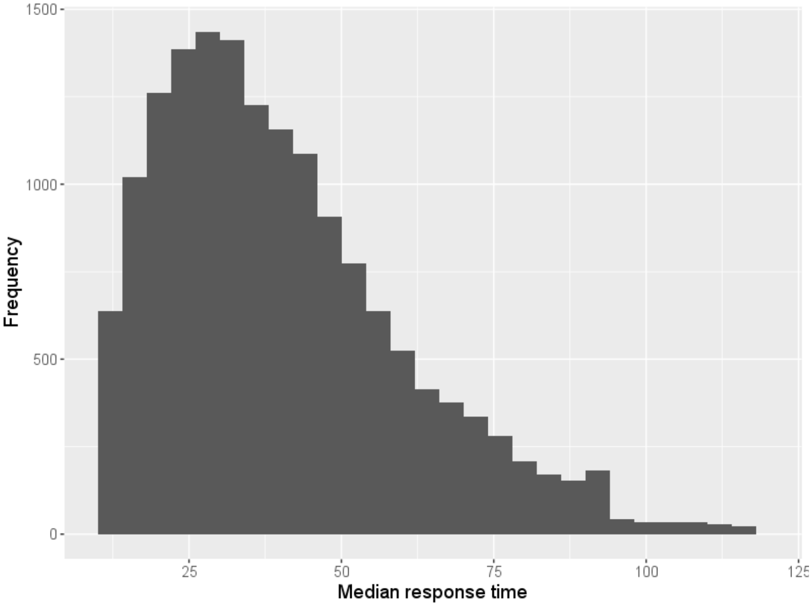


Figure 11:
Distribution of median response times of tests

The claim that SAM is a power test is further supported by looking at the distribution of individual response times. Figure 12 shows that the percentage of all responses are monotonically decreasing by the time spent on them until the end of the time limit where we find an increase (on the left side of the figure). This could be by both the automatic submission of responses at the end of the time limit or by test takers deciding to submit their responses in the last moments. This could indicate a time pressure on the test takers. However if we look at the percentage of correct answers (on the right of the figure) we see no increase at the last moments. We can also see that the correct answers submitted in the last moments remain below 1% of all the responses. This suggests that given more time the increase of correct answers would be negligible thus qualifying SAM as a power test. This calculation of course is unable to account for the effect of the known time limit – how much more time people *would* have spent on the tasks if there was no time limit at all.

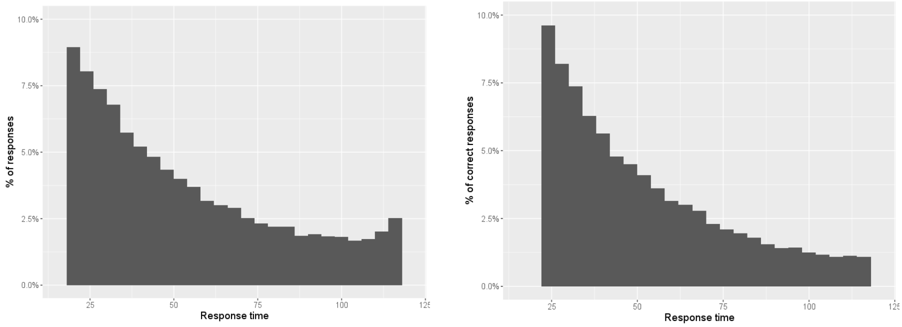


Figure 12:
Percentage of responses by response time

Performance and response time

When looking at the relation of performance and response time on an individual project level we usually see a strong (r = 0.56) positive correlation (Figure 13).

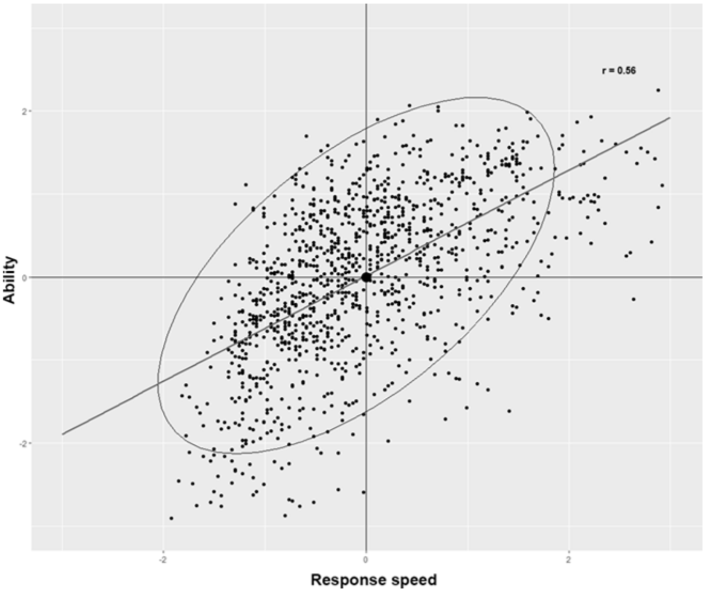


Figure 13:
Correlation between standardized performance and response time in the UNK project.

This suggests that more time spent on the tasks result in higher performance OR (since this is an adaptive test and to perform higher test takers need to answer more difficult items) that more difficult items require more time to solve. The difference between the two explanation is important. One of them suggests a motivational reason in the sense that those who are willing to spend more time with the test will have higher scores. The other explanation is a cognitive one claiming that although test takers get items that match their abilities the more difficult tasks take longer to solve – even for those with high ability. These explanations are of course not mutually exclusive and both effects are possible at the same time. This question requires further studies.

Looking at our results in the complete dataset (Figure 14) we can see that the relationship between time and performance is not linear. Spending more time with the test yields progressively less and less result and indeed spending more than 75 seconds with a task on average yields virtually no incremental performance increase.

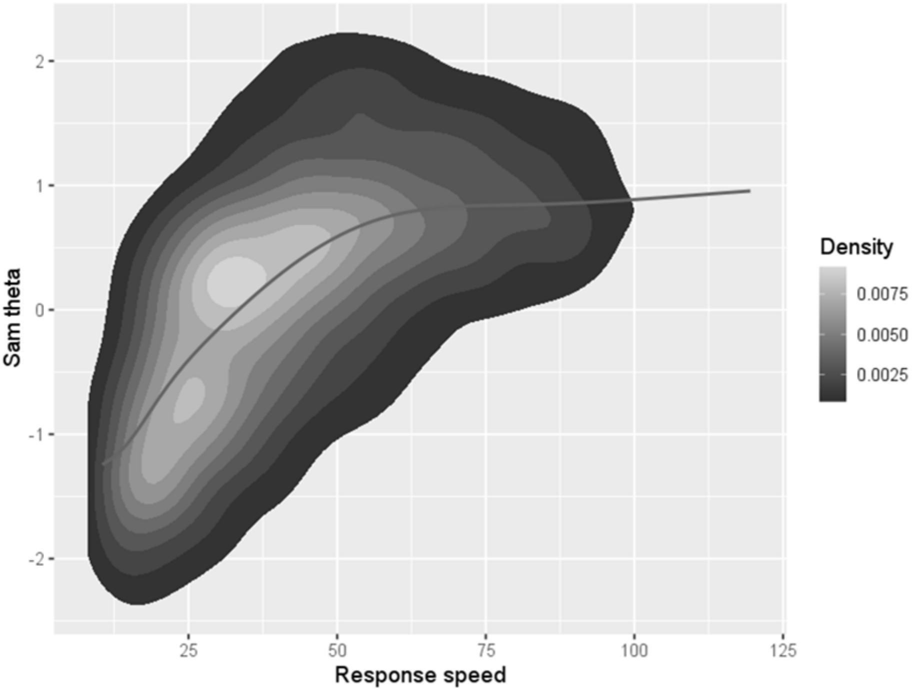


Figure 14:
Speed and performance in the complete dataset

Motivation and response time

On a group level test-takers in high-motivation projects spend significantly more time with the test even after factoring out the effect of higher performance (Figure 15). This suggests that there is an effect of motivation on time spent with the test beyond that of the more time required to solve more difficult tasks.

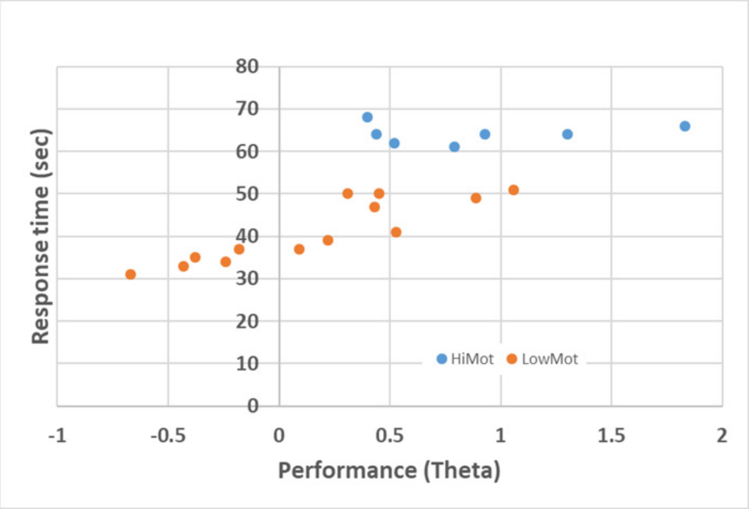


Figure 15:
Test performance and response time in high and low motivated test-taker groups

Further – non-conclusive – evidence of the effect of motivation to response time and performance came from the testing project of elementary school children where two online tests (SAM and a vocabulary scale called NOVO) were administered to school classes by a supervisor. The order of the two test was randomly determined for each class. The project was carried out in two phases identical to each-other except that in the first phase the children who finished earlier could play between the two tests and in the second phase they were asked to read on their own or stay silent. Figure 16 shows that in the first phase – where early finishers could play – the performance was significantly better and time spent on the tests were longer if they started with SAM and not with the Vocabulary scale; in the second project we found no such interaction. We know that response time and performance is related in the SAM test but not in the vocabulary. We believe therefore that the difference between the two phases are due to the motivational difference in the children – when early finishers could play they were less willing to devote more of their time and mental energy to cognitive problem solving. The same motivational difference does not affect the recollection of word meanings.

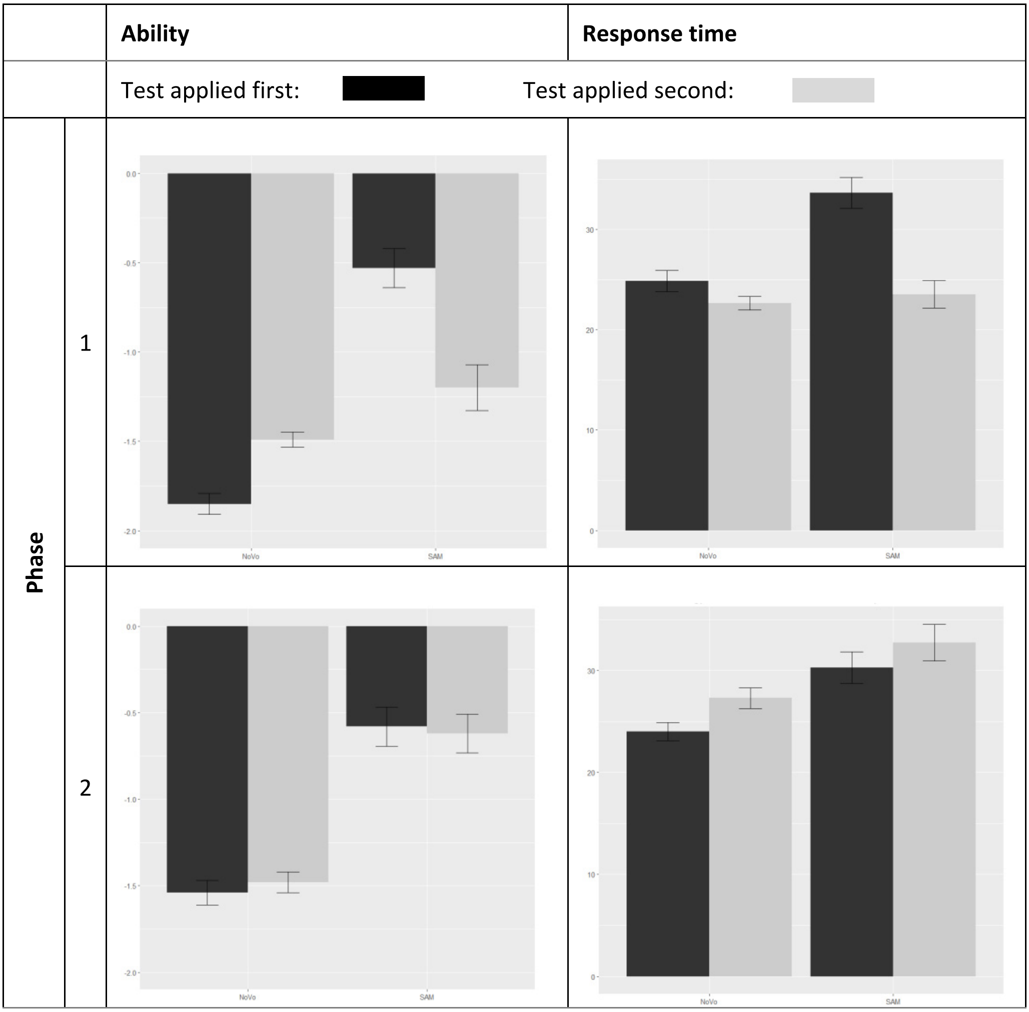


Figure 16:
Test performance and response time on SAM and NoVo by order and project conditions

The hypothesis that motivation plays an important role in response time is further supported by the fact that the ability level of the test taker and the difficulty of the item explains only 20% of the variation in response time.

The effectiveness of the adaptive algorithm

The adaptive algorithm aims to select items that would provide the most possible information about the ability of the test takers. Since ability follows a normal distribution an effective adaptive algorithm will use items with medium difficulty and high discrimination the most. Figure 17 shows that this is exactly the situation in our case. The group of items on the bottom of the chart are items currently trialed.

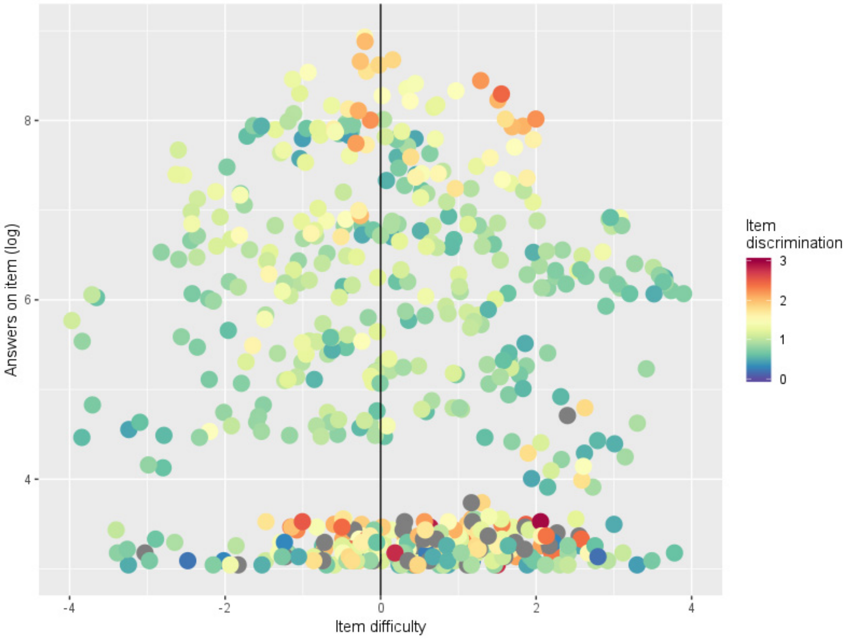


Figure 17:
Item use by item difficulty and discrimination

Next we will examine how effective the adaptive algorithm of the test proved to be in two ways. First in terms of how successful the algorithm was to generate different tests to prevent the leaking of the solutions. Subsequently we examine how well the algorithm could keep the difficulty of the displayed tasks at the ability level of the test taker, which is important for precise measurement as well as for keeping the test-takers motivated.

Item reuse (how different are two tests for the same ability level)

It is relatively easy to get metrics of item reuse from simulations. However – since there are so many external factors that can't easily be added to a simulation - we looked at our real-life data. Figure 18 shows the average number of common items (out of 25) between test takers of different ability levels. The image shows that test takers of identical ability levels get 10 common items on the average, while test takers with a difference of 3 only get 1 common item on the average. Of course the order in which these common items are displayed could still be very different. Therefore we conclude that at this point it seems to be difficult to obtain and disseminate a usable scoring key by filling out the test or – in a group testing setting – rely on the answers of other test takers.

	>1	1,0	0,-1	-1,-2	-2
-2					10
-1,-2				8	7
0,-1			10	7	4
1,0		10	7	4	2
>1	10	7	4	2	1

Figure 18:
Average number of common tasks by ability levels

Success rate of answers

In a well-balanced adaptive test we expect roughly half of the answers to be correct at every ability level. Figure 19 shows the percentage of correct answers in our projects by the average ability level of the participants in the project. Although there is a clear increase in the percent of the correct answers in absolute terms this has very little effect. While projects are in a very wide ability range (more than 2sd between the averages) the percent of correct answers remain in the range of 44%-63% - remarkably close to 50%. This shows that the adaptive algorithm can follow well the ability estimates and can provide optimal tasks to every ability level.

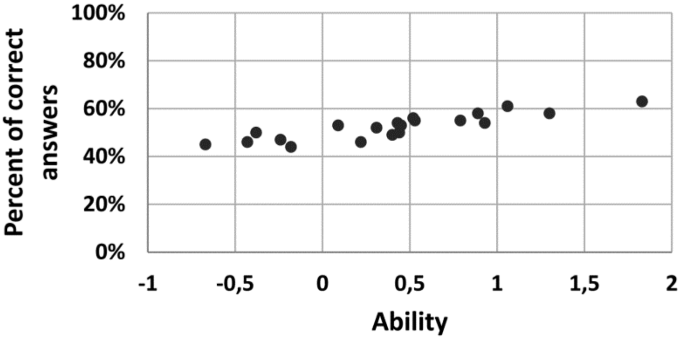


Figure 19:
Percent of correct answers by ability average in different projects

Test improvement and stability over time

One of the most important parameter in the quality of an adaptive test is the size and quality of its item bank. While in traditional testing the risk of the scoring key eventually leaking out increased with prolonged, widespread use in IRT based online testing it is possible to increase the quality proportionately with test use both by constantly adding new items and by refining the model with new data. The price of this continuous improvement is that we need to constantly monitor if results remain comparable between the recalculations of the model. In this section we present data showing that we can continuously improve the quality of the instrument by test use, while keeping the scores comparable over time.

Improvement

As new items and new response data are constantly added to the data bank and the model is periodically updated we expect a continuous improvement of the reliability of the test over time. Figure 20 shows how the – moving – average of the standard errors of the measurement changes over time. To counter the effect of different ability levels on the measurement error only tests with Theta estimate between -0.8 and 1.2 were used. The chart clearly shows that after the initial trial period (when items were still randomly displayed) the reliability of the test constantly improved. The most recent data shows that the average measurement error in this ability level is below 0.3 – which shows outstanding reliability.

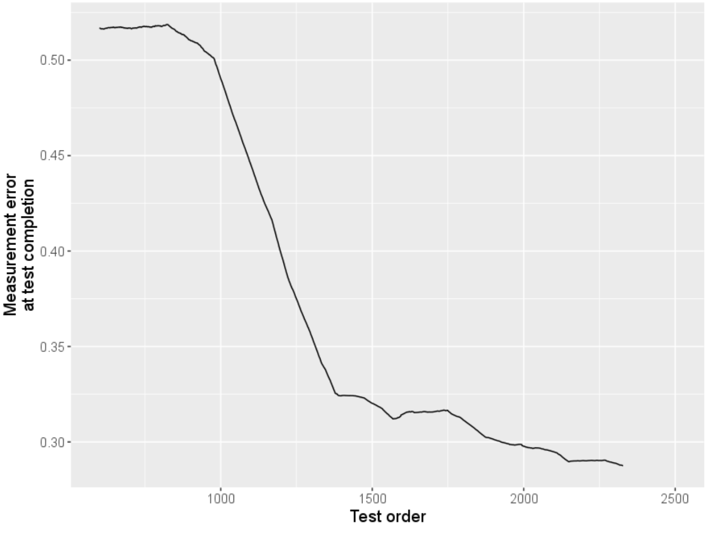


Figure 20:
Measurement error (SEM) around the mean by test order (recency)

Stability

It is important that scores remain comparable over time and despite of the changes in the item bank and estimation model. In this section we compare the estimates we made at the time of test completion with the estimate we would currently make using the all the currently available data. The timespan we can study our results this way is currently 2 years. Since the model has been regularly updated in the meantime we expect that recent estimates will be more similar to our current estimates than older ones.

Figure 21 shows the current and the original ability estimates for three projects taken at different times and with different average ability levels. The chart shows that the ability estimates stayed very stable over time and ability levels.

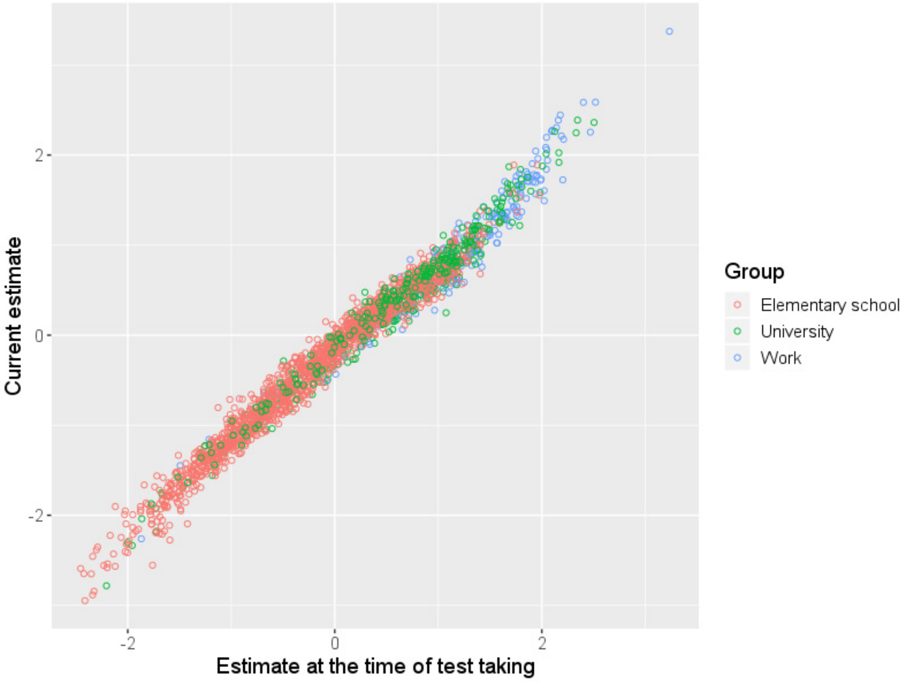


Figure 21:
Original and current ability estimates in different projects

To quantify stability correlation coefficients between current ability estimations and the estimations made at the time of finishing the test (with projects over 100 participants) were calculated. Not counting the two trial projects (where the discrepancies were expected) all other projects have a correlation higher than 0.96 showing that the ability estimations remain very stable over time despite of the changes in the item bank and the model (Figure 22).

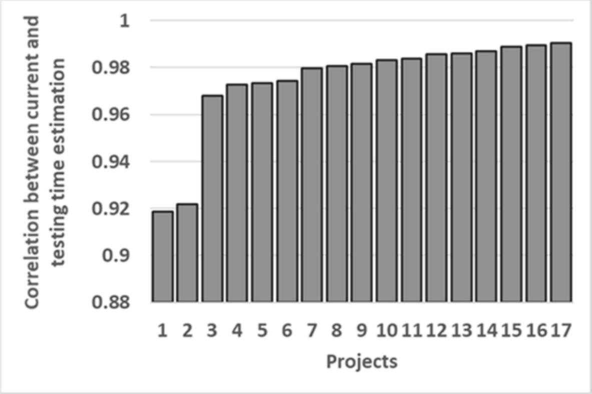


Figure 22:
Correlations between original and current ability estimations in different projects

Reliability

By using IRT based adaptive testing with a large and high-quality item bank we aim to achieve high levels of reliability with short tests. The continuous line on Figure 23 shows the theoretical reliability of the current item bank while the dots on the chart show the actual standard errors achieved during testing by using 25 tasks at different ability levels. The chart shows that between -2 and +2 the actual standard errors consistently remain well below 0.35 (the equivalent of close to 0.9 reliability coefficient) – which shows that our test indeed provides us with reliable results in a very wide ability range. The marginal reliability of the whole model is above 0.99.

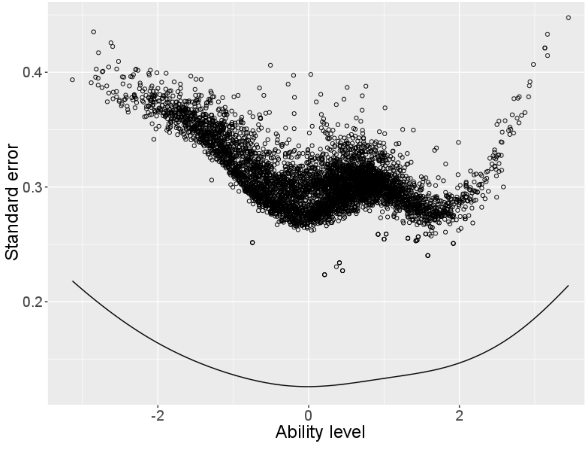


Figure 23:
Standard error of measurement by ability

Test-retest

Test-retest coefficient is a traditional measure of reliability. It must be noted that test-retest correlation figures are traditionally interpreted on fixed tests where test takers answered the same set of questions twice. This way one can ensure high test-retest correlation by simply answering the same set of questions the same way both times – independently from the construct reliability of the underlying instrument. In our adaptive setting the tasks are different in the two occasions (we previously showed that more than half of the items were previously unseen) so in this case any similarity between the results are mostly the result of the reliability of the test construct.

In our test-retest study we asked university students to do the SAM test 3 times in three different days but preferably within a week. Table 4 contains the correlations of the 3 testing occasions. Though these high correlations are – of course – highly significant, unfortunately the number of test takers in this project are yet very small, therefore we consider this to be only (very) promising preliminary results.

Table 4:
Test-retest correlations of SAM

	n	T1	T2	T3
T1	24	■		
T2	20	0.85	■	
T3	18	0.93	0.93	■

Learning effect

Using results from the test-retest study we can examine if and how much the performance of the test takers increased with practice. As literature predicted (Scharfen, Peters & Holling, 2018) there was a highly significant increase in performance between the first and the subsequent occasions but no significant difference between the second and the third occasion (Figure 24). When comparing the effect size it must be noted that retest effect between alternate forms (such as those generated by adaptive testing) are generally found to be significantly smaller than those between identical tests.

While performance increased response times dropped significantly with each subsequent testing occasions. On one hand it is expected that speed should increase with practice on the other hand though previously increasing performance was always linked to increasing response times. This finding requires further investigation.

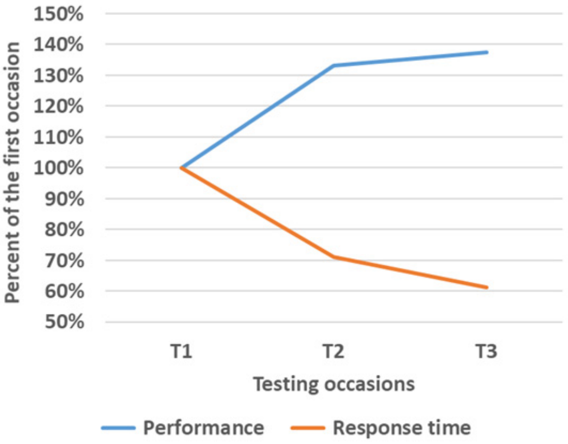


Figure 24:
Performance and response time by number of times the test is taken

Culture dependence

At this point we have limited data on culture dependence or independence. In the data we have we found no evidence of culture dependence on the test level. Figure 25 shows the standard error of measurement by ability for adults in Hungary and in China. In case of a

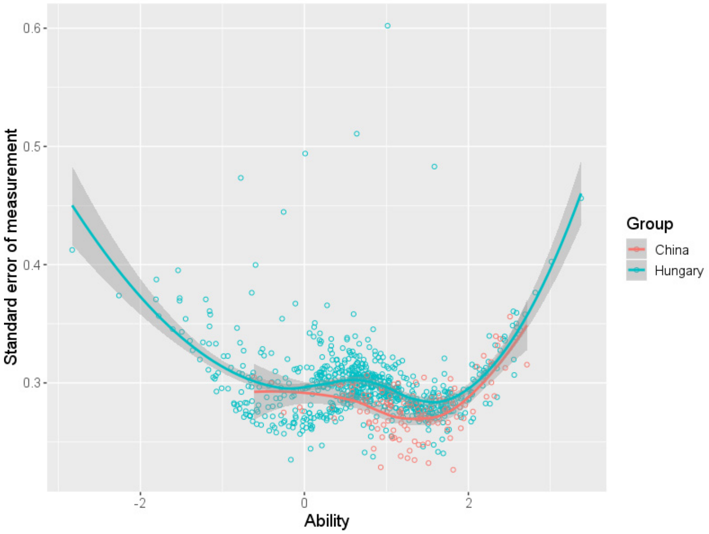


Figure 25:
Measurement error by ability level among Chinese and Hungarians

cultural bias we would expect to find higher level of error in the Chinese sample at the given ability levels than in the Hungarian sample. No such evidence was found - at every ability level the measurement errors are smaller in the Chinese sample than in the Hungarian sample. We are certainly aware of the fact that more studies are needed about the culture dependence of the instrument.

Validity

The validity of the instrument is the extent to which the instrument measures the intended construct. Out of the many existing validity types this study focuses on parallel validity results. In this section the relation of SAM with external variables such as other tests, performance measures or background variables are examined with the aim of finding hypothesized correlations.

Socio-economic status

As found in previous studies (Klein et al., 2008) there is strong positive correlation between the RPM and the socio-economic status (SES) of the test taker. We expected to find similar correlation between SES and SAM. It has been proved to be very difficult to obtain data about SES. On the individual level in open-access projects self-reported education of the test taker had a significant positive effect on test scores. On a group level elementary schools with higher SES had significantly higher average scores on SAM (r = 0.4, n = 60, p < 0.01). We found a significant difference between schools in different settlement types. In our near-representative study of 10 year old elementary school students schools in Budapest had significantly higher average SAM scores than schools in middle sized settlements that had still significantly higher scores than those in small villages.

Further studies are needed to establish the relation of SES and SAM scores on an individual level.

School grades

Cognitive abilities are important components of any learning process. Therefore we expected to find a significant correlation between SAM scores and school grades. Figure 26 shows a significant (r = 0.55) correlation between SAM results and school grades. There is an obvious ceiling effect on school grades suggesting that school grades do not differentiate between high performers in school and that this has a stronger effect on those with higher cognitive abilities. We found similar results in all our projects with elementary school children.

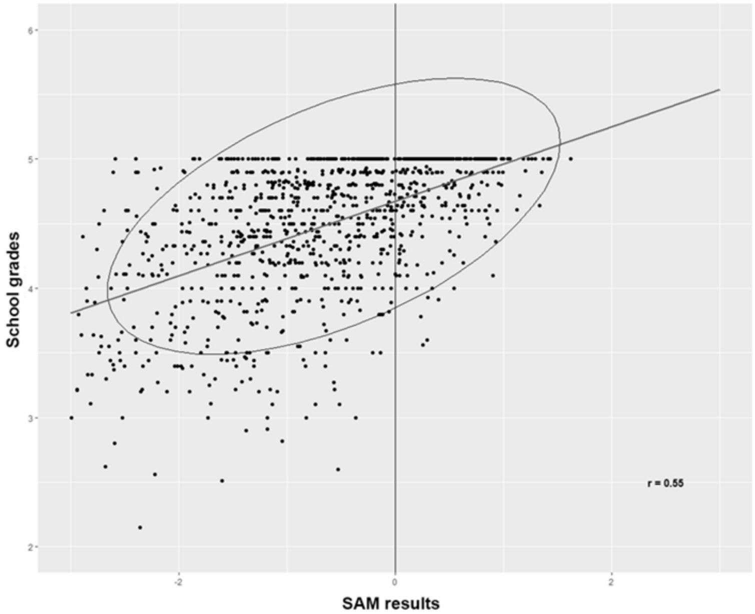


Figure 26:
School grades by SAM results

Standard Progressive Matrices

As the SAM aims to assess the same underlying construct than the RPM correlation with the Matrices tests are especially important figures of validity. Figure 27 shows results of Standard Progressive Matrices scores in relation to SAM scores of elementary school children. Correlation between the two tests were found to be r = 0.68 (n = 1060) which is somewhat lower than expected. Our primary explanation for this result is related to the restriction of range (participants were exclusively children around the age of 10) and the lack of motivation to complete the tests which could explain the larger discrepancies at the lower end of the results. A new project to assess the relation between the two tests are under way with university students. Having the finally established correlation between the two tests much lower than the test-retest correlations of the individual tests would raise important questions about the ways in which the tests differ. Further studies are needed to establish first a more reliable estimate of the correlation between the two tests.

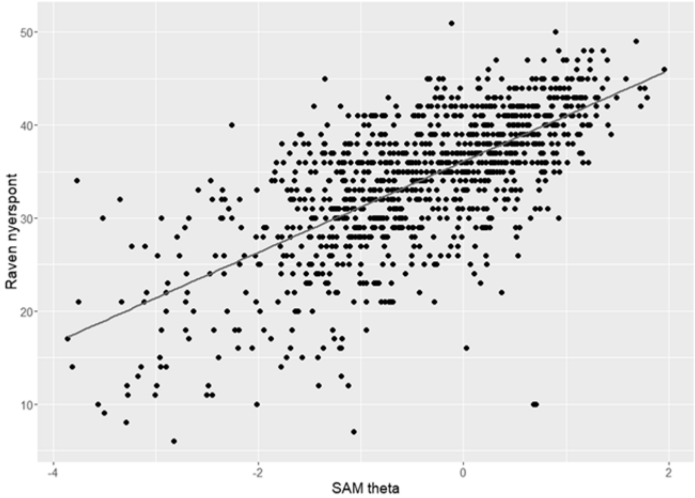


Figure 27:
SPM scores by SAM results

Competency measures

In different projects several competency measures were used with test takers together with SAM. In the following section the hypothesized correlations between these competency measures and SAM results are reviewed. Our hypothesis is that competencies related to task orientation and thinking will have high correlation with the SAM results in both self-descriptive and descriptive competency measures.

BETA Questionnaire

BETA Questionnaire is an ipsative, forced-choice, self-descriptive, online questionnaire for the world of work similar to DISC (Wallace, Clarke & Dry, 1956). Out of the 4 measured competencies "Patience" – working calmly and persistently while paying attention to tasks was found to correlate the highest with SAM with a small, but significant effect size (r = 0.18, n = 220). The effect of the ipsative question format must be taken into account when considering the relative low correlation size.

Spectrum Questionnaire

Spectrum Questionnaire is an ipsative, forced-choice, self-descriptive, online questionnaire measuring competencies with 360 questions from the world of work. Out of the 20 measured competencies "Effective thinker" – one who prefers building theories, making judgements, thinking quickly, thinking in systems and producing solutions – was found to correlate the highest with SAM with a small, but significant effect size (r = 0.15, n = 724). The effect of the ipsative question format must be taken into account when considering the relative low correlation size.

Renzulli Scales for Rating The Behavioral Characteristics of Superior Students

The Renzulli Scales is a normative questionnaire used by teachers describing the behavior of students (Renzulli et al., 2002). Out of the 14 measured scales Motivation, Learning, Mathematics and Planning had the highest correlations with SAM results (r between 0.5-0.45, n = 500) while music had the lowest (r = 0.1, n = 500).

Proprietary questionnaire for teachers to rate students

This questionnaire was created as a trial instrument and was used in the second phase of the same project than the Renzulli scales were used for similar purpose. The questionnaire yielded information about the children described by the teachers on 7 scales. Mathematics was found to correlate highest with SAM results (r = 0.48, n = 183) while Motion was found to relate the least (r = 0.1, n = 183).

Altogether from the correlations collected between competency questionnaire data and SAM results we can conclude that both according to self-description data and teacher ratings SAM measures a behavioral feature in the domain of thinking, learning and mathematics.

Score differences

Ability scores in SAM are calculated using IRT estimation and given in Theta (θ) which has a theoretic standard normal distribution. Figure 28 displays the frequencies of actual ability estimates showing that the distribution of actual results follow closely the theoretic standard normal distribution.

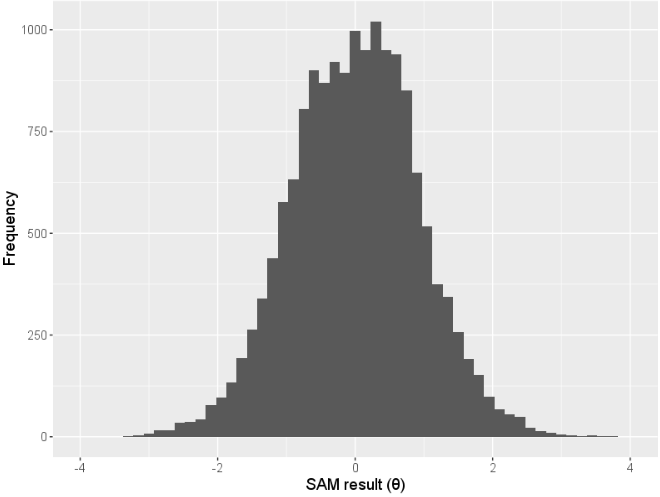


Figure 28:
Distribution of SAM results

Age

One of the fundamentals of developmental psychology is that abilities increase in early childhood. Therefore we expect to see an increase in ability scores in the early years – though we have no clear hypothesis until what age should this increase last. Figure 29 shows the average SAM results by the age of the test taker. From the figure we can clearly see a significant and large increase of performance in the 9-15 age range (n = 3738) . Data in later years start to be more sparse and unreliable – greatly affected by the different samples taken.

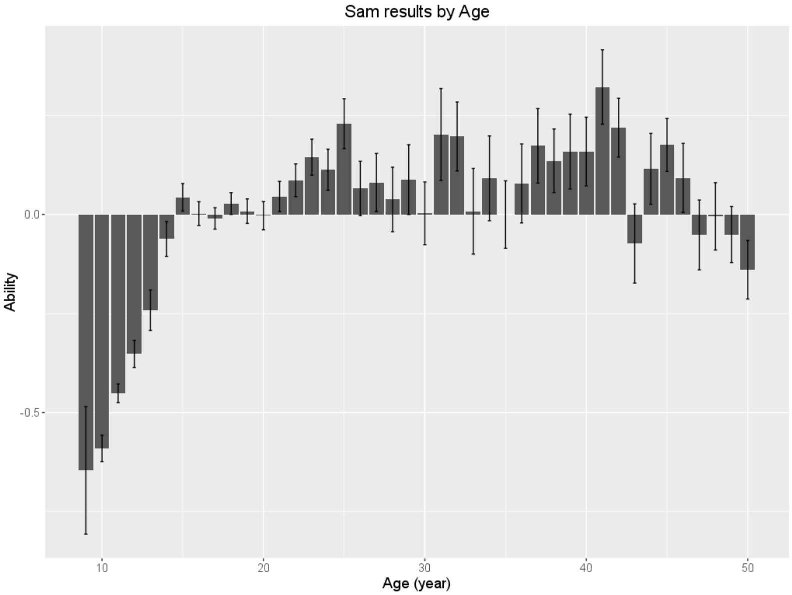


Figure 29:
Average SAM results by age

Though there is a large increase in performance by age in the early teenage (9-15) years it must be noted that there are large individual performance overlaps between age groups. Figure 30 shows the distribution of performance in the early years with the large overlaps between the groups.

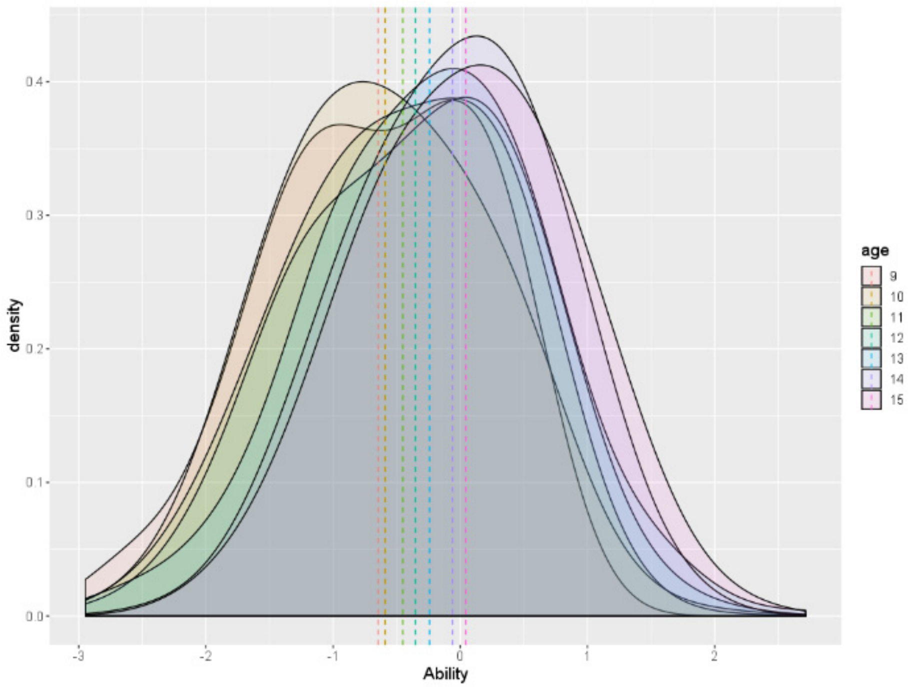


Figure 30:
Ability distributions between 9-15 years of age

Gender

In the closed projects of our dataset two significant differences between the performance of males and females were found – one of them showed higher results for males, the other for females. Altogether there was no evidence of gender difference on test performance. Open projects were excluded from the analysis because of the unreliable nature of participation criteria (e.g. the Facebook algorithm in our experience tended to reach significantly more women than man).

Limitations

Current data suggests that SAM results explain around 50% of the variations of the RPM results. Future research is needed to establish to what extent are the two test measuring different things (e.g. the study of those who get high scores on one but low on the other should shed some interesting light on this).

Further research is also needed to set up age norms for all ages and to examine the developmental tendencies of eductive ability. It is also recommended to work on the validity issues and compare SAM results to various other ability tests and nonperformance assessment measures.

Imposing a time limit on a power test always corrupts its purpose to some extent. The online format of the test enables us now to easily collect and study response time. Using this data attempts have been made in the current study to establish the effect time limit has on the results. In general the study of the relationship between the time spent to come up with a solution, item difficulty and motivation is probably one of the most interesting aspects of the current study.

Cultural differences in the test are also a challenging and interesting topic for future research.

Summary

In the present study we provided evidence to the validity and reliability of the SAM test. We showed that this new online adaptive test of eductive ability measures quickly and reliably the cognitive construct it aims at in an extremely wide ability range which enables its use with children and adults alike in both supervised and unsupervised manner. We believe that overall this test is an exciting new development that takes forward the idea of measuring the ability to make sense out of confusion in situations where information is inherently misleading or controversial. Science – and life in general – is full of such situations.

References

Cattell, R. (1987). *Intelligence (Its structure, growth, and action).* Advances in Psychology, 35. North Holland, p. 693. (eBook ISBN: 9780080866895)

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. In: *Encyclopedia of Human Intelligence* (Ed. R. J. Sternberg). New York: Macmillan, pp. 443-451.

Kaplan, R., Saccuzzo, D. (2012). *Psychological testing: Principles, applications, and issues.* Cengage Learning.

Klein B., Klein S., Joubert K. & Gyenis Gy. (2008). Social Cage (socio-economic status and intelligence in Hungary). In: Raven, John-Raven, Jean (Eds.): *Uses and Abuses of Intelligence (Studies Advancing Spearman and Raven's Quest for Non-Arbitrany Metrics).* Royal Fireworks Press-Competency Motivation Project Edinburgh, Edge 2000 Ltd. Budapest, RTS Romanian Psychological Testing Services, Bucarest, 568-593.

Neisser, U. (1976). *Cognition and reality. Principles and implication of cognitive psychology.* San Francisko: WH Freeman and Company. Available from: https://www.researchgate.net/Ulric-Neissers-Perceptual-Cycle-Source-Redrawn-from-Ulric-Neissers-Cognition-and_fig1_300015484

Pearson (2018) https://www.pearsonclinical.com/psychology/products/100000414/ravens-advanced-progressive-matrices-apm.html

Raven, J. (2008 a). Stability and change in norms over time and culture: The story at the turn of the century. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics.* Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project; Budapest, Hungary: EDGE 2000; Cluj Napoca, Romania: Romanian Psychological Testing Services SRL. (Chapter 8, pp. 213-257). http://eyeonsociety.co.uk/resources/UAIChapter8.pdf

Raven, J. (2008 b). General introduction and overview: The Raven Progressive Matrices Tests: Their theoretical basis and measurement model. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics.* Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project; Budapest, Hungary: EDGE 2000; Cluj Napoca, Romania: Romanian Psychological Testing Services SRL. (Chapter 1, pp. 17-68). Also available at http://eyeonsociety.co.uk/resources/UAIChapter1.pdf

Raven, J. C. (1936). *Mental tests used in genetic studies. (The performance of related individuals a tests mainly eductive and mainly reproductive)*. Unpublished Master's Thesis, University of London.

Raven, J., & Fugard, A. (2008). *What's wrong with factor-analyzing tests conforming to the requirements of Item Response Theory?* WebPsychEmpiricist, May 23. http://wpe.info/papers_table.html or http://eyeonsociety.co.uk/resources/fairttsts.pdf

Raven, J., Prieler, J. & Benesch, M. (2008). Using the Romanian data to replicate the IRT-based Item Analysis of the SPM+: Striking achievements, pitfalls, and lessons. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics.* Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project; Budapest, Hungary: EDGE 2000; Cluj Napoca, Romania: Romanian Psychological Testing Services SRL. (Chapter 5, pp. 127-159). Also available at: http://www.wpe.info/papers_table.html

Renzulli, J. S., Westberg, K. L., Hartman, R. K., Callahan, C. M., White, A. J., & Smith, L. H. (2002). *Scales for Rating the Behavioral Characteristics of Superior Students. Technical and Administration Manual.* Revised Edition. Creative Learning Press, Incorporated.

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence, 67*, 44-66.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. *Advances in the psychology of human intelligence, 2*(S 47), 103.

Spearman, C. (1924). The nature of "intelligence" and the principles of cognition. *Journal of Philosophy 21* (11), 294-301.

Spearman, C. B. (1927). *The Abilities of Man (Their Nature and Measurement)*. New York, Macmillen, pp. 221. (The Blackburn Press, 2005)

Taylor, N. (2008). Raven's Standard and Advanced Progressive Matrices among adults in South Africa. In J. Raven & J. Raven (Eds.) *Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics.* Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project;

Budapest, Hungary: EDGE 2000; Cluj Napoca, Romania: Romanian Psychological Testing Services SRL. (Chapter 15, pp. 371-391). http://eyeonsociety.co.uk/resources/UAIChapter15.pdf

van Alphen, A., Halfens, R., Hasman, A. & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing, 20*, 196-201.

Wallace, S. R., Clarke, W. V. & Dry, R. J. (1956), The Activity Vector Analysis as a Selector of Life Insurance Salesmen. *Personnel Psychology, 9,* 337-345. doi:10.1111/j.1744-6570.1956.tb01072.x

Web (2018). A few of the websites currently offering RPM-like tests:
https://iqpro.org/
https://iq-research.info/en
https://www.highiqpro.com/iq-tests-iq-scores-iq-questions/matrix-iq-brain-teasers
https://iqhaven.com/mytests1/matrix-g_test/thetest.htm
https://testyourself.psychtests.com
http://www.iqtesztek.hu/hu-hu/iqtest/1

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.