Transitionen von der Erstausbildung ins Erwerbsleben
Transitions de l'Ecole à l'Emploi
Transitions from Education to Employment

UNIVERSITÄT
BERN

TREE TECHNICAL PAPERS SERIES NO. 4

IMPLEMENTATION OF A COGNITIVE ABILITY TEST IN THE TREE2 PANEL SURVEY

DOMINIQUE KREBS-OESCH
BEN JANN
SANDRA HUPKA-BRUNNER

Bern, 2023

source: https://doi.org/10.48350/183109 | downloaded: 10.5.2024

Impressum

Series edited by TREE (Transitions from Education to Employment).
University of Bern
Fabrikstr. 8
3012 Bern/Switzerland

www.tree.unibe.ch

tree.soz@unibe.ch

Suggested citation

Krebs-Oesch, D., Jann, B., Hupka-Brunner, S. (2023). Implementation of a cognitive ability test in the TREE2 panel survey. TREE Technical Paper Series No. 4. Bern: TREE. doi 10.48350/183109

This work is licensed under a Creative Commons Attribution 4.0 International License.
https://creativecommons.org/licenses/by/4.0/legalcode

Table of contents

1 Introduction 4

2 What does the test measure? 5

3 Implementation in the TREE2 panel survey 6

4 Data 9

5 Empirical analyses 9

5.1 Formal validity of the test scores 9

5.2 Descriptives 10

5.3 Assessing test functionality and validity in TREE2 12

6 Conclusion and some words of caution 15

References 16

Appendix: Overview of available para-data 17

1 Introduction

The focus of the present document is on a test of basic cognitive abilities implemented in the panel survey of the second TREE cohort (TREE2). Basic cognitive abilities are crucial for academic and professional success, as they form the basis for effective learning and information processing. The inclusion of the test in the TREE2 panel survey hence provides a database allowing to investigate a wide range of interdependencies between different abilities (mathematics skills, basic cognitive abilities and reading speed; for more detail on TREE2's test design, see Hupka-Brunner et al., 2023) educational pathways and the life course in general.

To this end, TREE has drawn on the KFT (*Kognitiver Fähigkeitstest*) as developed by Heller and Perleth (2000), to which we will henceforth refer as CAT (cognitive abilities test). Heller and Perleth, on their part, draw on a cognitive abilities test originally conceived by Thorndike and Hagen (1971, 1993, in: Heller & Perleth, 2000), which they have translated to German and adapted for educational test contexts. We have implemented an online version of the German CAT's subtest N2 focussing on figural analogies. The nonverbal subtest was administered in the baseline survey to one split-half of the TREE2 sample[1].

We designed the online version of the test for laptop or computer administration, but the test could also be completed on smartphones or tablets. By adapting the paper-and-pencil instrument to a digitised web format, we were able to extend the test's use beyond the proctored classroom setting that the original format was designed for. Being able to conduct the test outside the classroom and without a test administrator furthermore greatly facilitates test administration and data collection. Moreover, the adaption to web-based administration provides valuable process data on respondents' test behaviour. Such para-data allow, e.g., to draw conclusions with respect to the formal validity of individuals' test results. Although the test itself is non-verbal, we took great care to adapt test instructions to the self-administered web-based setting and to translate it to TREE's other two survey languages (French and Italian) in order to have full test coverage of our sample.

The document at hand intends to inform on conceptional considerations as well as on the details of the adaptations that we applied to Heller and Perleth's original test. We further provide results with respect to the test's reliability, criterion validity and some words of caution for scholars who wish to analyse the test data.

[1] For more detail on TREE2's survey and sampling design, see Section 3 and Hupka-Brunner et al. (2023).

2 What does the test measure?

"A very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience." (Gottfredson, 1997)

Beyond Gottfredson's definition of intelligence, theoretical approaches to conceptualizing intelligence have always been inconsistent. Brunner et al. (2014) point to an important distinction for cognitive theories across the lifespan, namely the differentiation between biologically vs. culturally determined components of intelligence. Two-component theories such as those of Cattell and Horn (Cattell, 1963; Horn & Noll, 1997), which distinguish between fluid and crystallised intelligence, reflect this differentiation. Accordingly, acquired knowledge is part of crystallised intelligence, whereas advanced cognitive processes such as reasoning (speed, accuracy, and coordination of cognitive processes) are attributed to fluid intelligence. Brunner et al. (2014) hold that experts consider fluid intelligence as a core aspect of the intelligence concept. With its sub-components processing speed and reasoning, it is one of the most important psychological constructs and is predictive for many learning processes, health and the achievement of various goals throughout one's life course.

The Cognitive Ability Test is a differential intelligence test designed to assess the cognitive capabilities of students from 4th to 12th grade. It provides information on linguistic, quantitative and nonverbal-figural thinking, including aspects of spatial thinking and the overall cognitive performance level of a student. The test is particularly suitable for educational and career counselling. By virtue of setting time limits for individual subtests, the test can be classified as a combined power-speed test. The figural or nonverbal subtest N2 of *Kognitiver Fähigkeitstest* (KFT, see Heller & Perleth, 2000) that TREE adopted measures reasoning, is suitable for administration from the age of 7 and includes grade-specific tasks for German grade levels 4 to 12. The subtest N2 comprises 25 test items of the type displayed in Figure 1. Respondents are granted eight minutes to complete the 25 items.[2]

Each task of the implemented subtest starts with a pair of figures or drawings that share a logical relationship. The respondents' task then consists of determining the relationship (analogy) between the two figures. There is always a third figure in the assignments, representing the first figure of a second pair. There are five figures to choose from for the second figure of the second pair. Respondents are called to find the one that matches the first figure in the same way as with the first pair.

[2] Net test solving time, excluding the introduction. Note that we implemented only one of the two (equivalent) test booklets developed for the original paper-and-pencil test (form A, see Heller & Perleth, 2000).

The test was developed more than twenty years ago. A revision and restandardisation is currently being conducted at the University of Rostock. For details see https://www.ipprdk.uni-rostock.de/forschung/diagnostik-und-forschungsmethoden/revision-des-kognitiven-faehigkeits-tests-kft/.

For reasons of copyright and confidentiality, we are not allowed to divulge items of the administered test. However, the following example illustrates the principle of the test with letters instead of figures.

Figure 1: Logical structure of CAT test items

	□	□	□	□	□
A : a = B :	*B*	*c*	*b*	*d*	*D*

3 Implementation in the TREE2 panel survey

Several criteria were considered to assess basic cognitive abilities in the TREE2 panel study. First, due to the presence of multiple national languages, a non-verbal, figural test was given preference to ensure comparability across Switzerland's various language regions (German-, French- and Italian-speaking). Second, a time-efficient test was required, as TREE is mainly designed to a comprehensive and detailed longitudinal capture of respondents' educational and labour market trajectories. The test therefore had to be short so as not to increase overall survey burden. Moreover, cognitive ability tests are often designed for individual diagnosis and may not be suitable for large-scale assessments.[3]

Consequently, the choice fell on the figural subtest N2 of the KFT 4-12, R, which has been utilized in various large-scale assessments, including PISA 2000/2009, ELEMENT, PALMA, NEPS, and IQB national comparisons, where it primarily functions as a control variable (Brunner et al., 2014; Scharenberg, 2012). Thus, the choice fell on a proven test that should allow comparisons with other Large Scales Assessments (see Section 5.3).

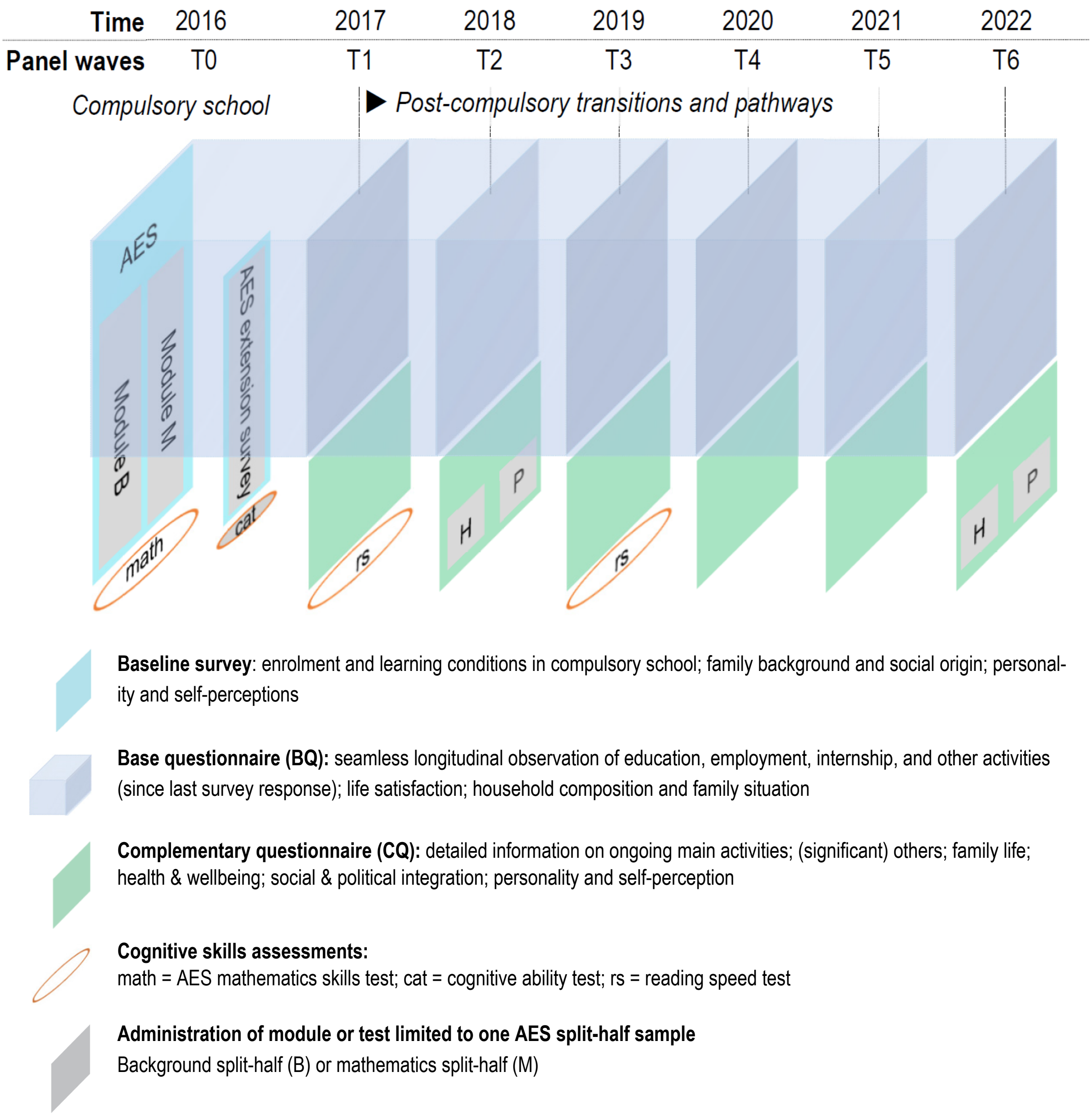
The test was administered to one split-half of the initial panel sample in the context of an extension to the baseline survey. Although both split-halves had been tested in mathematics at baseline, their student background questionnaires differed in content. The split-half in question provided comprehensive information on self-concepts and attitudes with respect to learning in general and mathematics in particular (module M in Figure 2; for more detail on the sampling design, see Figure 2 and Hupka-Brunner et al., 2023).[4] The extension survey with the

[3] Cognitive ability tests designed for individual diagnosis may not be suitable for large-scale assessments due to factors such as:
Length: They may be overly time-consuming.
Administration: They often require one-on-one administration by trained professionals, which may not be feasible or fundable in large-scale settings.
Interpretation: Their results may need expert interpretation, whereas large-scale assessments usually yield standardised scores that can be easily compared and analysed.
Content: They often focus on specific cognitive domains or abilities that are not relevant or necessary for large-scale assessments, which usually aim to measure general cognitive abilities or common domains across participants.

[4] The other split-half (module B) was administered a student background questionnaire that was co-developed by TREE, designed to collect information on a wide range of respondents' resources, their families and the schools they were attending

CAT test was conducted shortly (i.e., few weeks) after the main baseline assessment and may thus be considered as a synchronous measure with respect to the mathematics assessment.[5]

Figure 2: TREE2 panel design (up to wave 6)



Baseline survey: enrolment and learning conditions in compulsory school; family background and social origin; personality and self-perceptions

Base questionnaire (BQ): seamless longitudinal observation of education, employment, internship, and other activities (since last survey response); life satisfaction; household composition and family situation

Complementary questionnaire (CQ): detailed information on ongoing main activities; (significant) others; family life; health & wellbeing; social & political integration; personality and self-perception

Cognitive skills assessments:
math = AES mathematics skills test; cat = cognitive ability test; rs = reading speed test

Administration of module or test limited to one AES split-half sample
Background split-half (B) or mathematics split-half (M)

The research team collaborated with the test editor Hogrefe (Beltz Test GmbH) to adapt and revise the N2 subtest (Figure Analogies) of the CAT for CAWI administration in the TREE2 study. This involved several steps to ensure that the test be appropriate for the target audience and the web-based mode of administration:

at the time of the survey. Apart from the administration of the CAT test, the extension survey at baseline served to complete this information for the "M" split-half as well.

[5] The average lag between baseline and extension survey was approximately one month.

1. *Web adaptation*: The original paper-based test was adapted for web-based use. This involved displaying two test items (2 figure analogies) per screen and implementing a "Next" button to navigate between the 13 screens of task pairs[6]. Multiple checking was allowed in order to replicate the setting of the original paper-based test.

2. *Introduction*: The original introductory section, which explained how to read and understand the figure analogies, was adapted for the web version. Animated graphics were employed to enhance the understanding of the explanations provided. In the original paper version, the introduction lasted approximately four minutes and was read by the test administrator using an overhead projector. Some parts of the figure analogies are initially covered and then gradually revealed for better comprehension. This procedure is crucial to understand the figure analogies, so we paid close attention to replicate it in the web adaptation by initially hiding certain parts of the images and gradually revealing them as respondents click to advance, providing additional explanatory text in the process.

3. *Linguistic simplifications & translations*: The original introductory text, edited for test administration in Germany in 2000, was adjusted to suit Swiss-German vocabulary in consultation with Hogrefe (Beltz Test GmbH). As there is no test administrator available in the web variant to answer questions, we simplified the wording of the instructions in order to ensure respondents' comprehension. The adapted German version was then translated into French and Italian so that the test could be administered to TREE respondents from all language regions.

4. *Support*: It is important to note that the test assignments were not self-explanatory, and the introductory text with examples is relatively complex. In the web adaptation, a hotline and an email address was therefore provided for participants to ask questions or seek clarification. We thus aimed at emulating the setting of the test's proctored classroom administration. However, none of the respondents took advantage of this support function.

The revisions and adaptations listed above were adopted to ensure that the test be suitable for the TREE2 study's target sample and for web-based administration while maintaining the integrity of its original. For a better understanding of the adjustments, we also provide para-data from the web-based test, which can comprise additional information about the formal validity of the completed test assignments.

Unfortunately, we do not have any information about the devices that respondents used when they completed the test. We strongly recommended completing the survey on a desktop or laptop computer. However, we do not know if and how many participants attempted to take the test on a smartphone. As the test's technical implementation was not device-adaptive,

[6] The final screen contains only one analogy. This 2-item-per-screen setup was specified by Beltz Test GmbH.

navigating through the test on a smartphone was significantly more difficult, potentially leading to longer test times and possibly lower test scores.

4 Data

We achieved a total of 3341 valid tests, for which test scores were generated accordingly. Ten respondents encountered technical difficulties when completing the test. 266 respondents (approximately 6%) refused to take the test or terminated it prematurely. In four cases, the response patterns lead us assume that the test was not taken seriously. Furthermore, 204 respondents (about 5%) had difficulty understanding the test instructions. They either commented on their lack of understanding or were unable to complete a task correctly.

In the 2023 TREE2 data release (TREE, 2023), the CAT's data and para-data are provided in a specific dataset ('*TREE2_Data_Wave_0_cat_v2*'). This dataset not only includes the sum score of the correctly solved test items (variable *cat_score*) but also para-data for score validation (*cat_status, cat_rpattern, cat_timeisup, cat_lastpage, cat_comment*; see also Appendix) and the 25 test items (*cat_item_1-25*). These para-data provide detailed information on respondents' problem-solving behaviour. We provide a script (in Stata format) that reveals how we validated the test results and generated the sum score ('*TREE2_Syntax_Wave_0_CAT_Validation_v2*'). This script may be modified for individual use if needed. Additionally, the CAT sum score (*cat_score*) is also included in the general wave-specific dataset for panel wave 0 (baseline).

5 Empirical analyses

5.1 Formal validity of the test scores

A sum score was calculated for valid tests only. No score was computed if respondents…:

1. …did not complete the test (test break-off).
2. …displayed a pattern of response behaviour that indicates random or non-serious engagement[7] with the test.
3. …encountered technical problems during the test administration that affected their ability to complete the tasks accurately (mostly identified through respondents' comments).
4. …commented that they did not understand the test instructions or were unable to solve an item.

By excluding these cases, the sum score reflects a more accurate representation of the participants' cognitive abilities as measured by the test.

[7] E.g., always checking the first answer category (non-serious response pattern).

5.2 Descriptives

In the following, we provide some descriptive statistics of the data on test results as published in the 2023 data release (‘*TREE2_Data_Wave_0_CAT_v2*’, see TREE, 2023). Table 1 displays the frequencies of the sum scores. Information on para-data is displayed in the Appendix. Figure 3 visualises the distribution of test scores by the level of academic requirements of the lower-secondary programme attended, while Figure 4 does the same for the language regions. As the graphs reveal, the distribution of the test scores is either mostly right-skewed, in line with the findings in the manual of the original test[8] (Heller and Perleth 2000, p. 18). Overall, we observe a distinct bimodal pattern in the distribution of the TREE2 test scores. Among the students attending lower-secondary programmes with low academic requirements and contrary to the overall pattern, the distribution curve is left-skewed. The same is the case for the Italian-speaking subsample.

Table 1: Frequencies of CAT sum score

cat_score

		Freq.	Percent	Valid	Cum.
Valid	1	82	2.14	2.45	2.45
	2	94	2.46	2.81	5.27
	3	101	2.64	3.02	8.29
	4	125	3.27	3.74	12.03
	5	113	2.95	3.38	15.41
	6	118	3.08	3.53	18.95
	7	98	2.56	2.93	21.88
	8	72	1.88	2.16	24.03
	9	88	2.30	2.63	26.67
	10	64	1.67	1.92	28.58
	11	75	1.96	2.24	30.83
	12	80	2.09	2.39	33.22
	13	86	2.25	2.57	35.80
	14	94	2.46	2.81	38.61
	15	123	3.22	3.68	42.29
	16	155	4.05	4.64	46.93
	17	162	4.24	4.85	51.78
	18	192	5.02	5.75	57.53
	19	228	5.96	6.82	64.35
	20	236	6.17	7.06	71.42
	21	230	6.01	6.88	78.30
	22	239	6.25	7.15	85.45
	23	230	6.01	6.88	92.34
	24	175	4.58	5.24	97.58
	25	81	2.12	2.42	100.00
	Total	3341	87.35	100.00	
Missing	.v Invalid answer [Random/non-serious response pattern]	4	0.10		
	.w Invalid [Did not understand test instructions]	204	5.33		
	.x Invalid [Technical problem]	10	0.26		
	.y Invalid [Break-off, no answer]	266	6.95		
	Total	484	12.65		
Total		3825	100.00		

[8] Heller & Perleth (2000) indicate that the slightly skewed distribution of overall performance does not affect the ability to differentiate overall performance of the CAT N2 Subtest.

Figure 3: Distribution of sum scores by type of lower-secondary programme attended

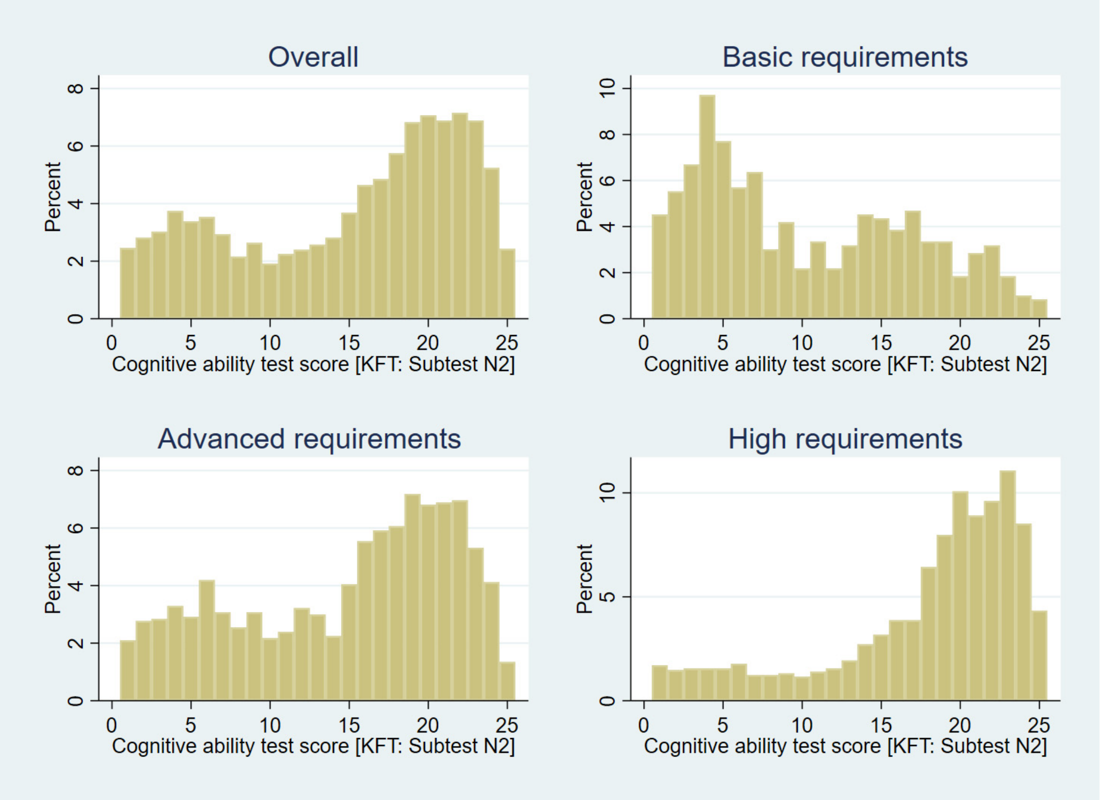
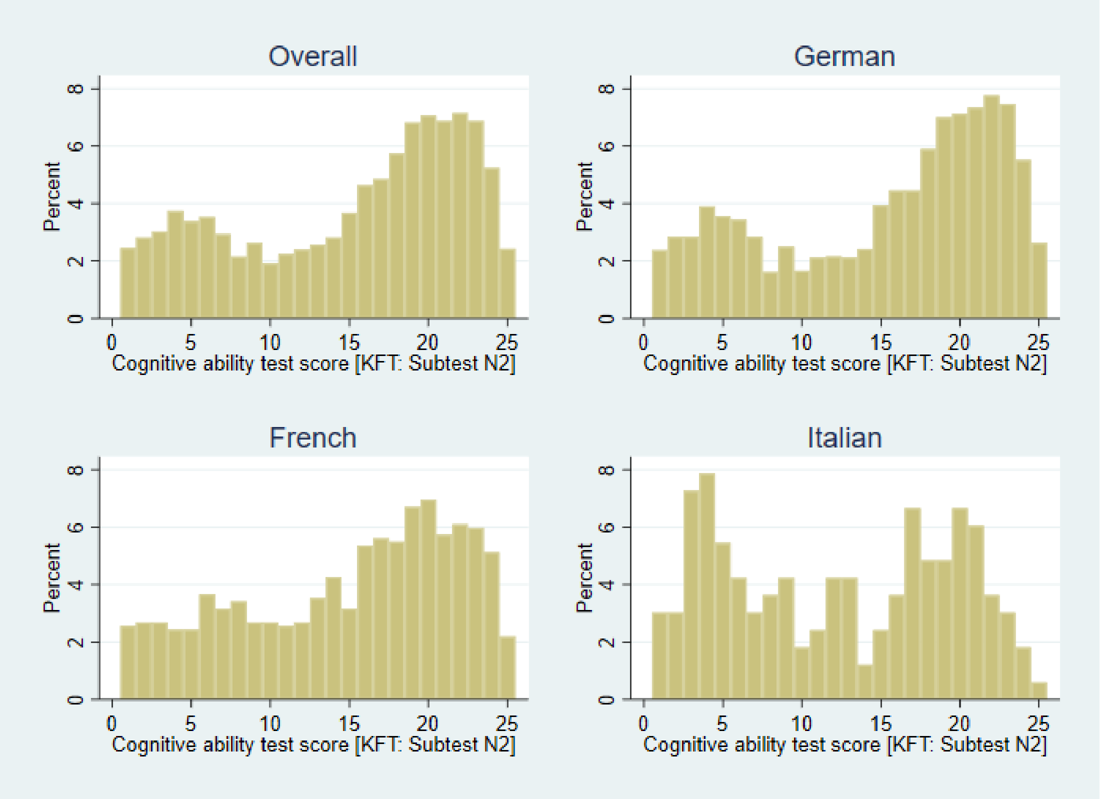


Figure 4: Distribution of sum scores by language region



5.3 Assessing test functionality and validity in TREE2

Since we could not conduct a paper-and-pencil test among a control sample, assessing the validity of the data is not straightforward. Since the norms of the original test were established for the paper-and-pencil version, they cannot be indiscriminately inferred to a computer version, especially in view of the time limit set for the completion of the test.[9] We are therefore in need of similar studies that have administered the test in order to compare the characteristics and distribution of the test scores. Furthermore, we focus on construct validity to examine the extent to which the test confirms hypothesised correlations.

Factors that may have influenced the test results include the test instructions, which are quite demanding. Within a classroom setting, test administrators can easily clarify questions brought up by the students to be tested. The online instruction that we developed strove to visualize and explain the test assignment as well as possible. However, the numerous comments made by the respondents suggest that a significant share of them failed to understand the instruction – including an unknown share of those who did not leave a comment. Moreover, the high share of low test scores in the Italian-speaking region suggests that the Italian translation seems to have been particularly demanding. A review of the data indicates that 14% of the Italian-speaking sample either discontinued the test or left a comment suggesting difficulties with the introduction. A similar pattern is observed in the French version of the test (15%). The German version triggers a slightly lower percentage of discontinuations and comments (11%).

The instructions do not explicitly state that only one solution is correct, i.e., that multiple answers are considered incorrect. Since a considerable proportion of respondents provided both single and multiple answers (1227 cases; 32% checked several answers on one or several occasions), some test scores are very low. This calculation is in line with the official manual. However, it is suspected that multiple answers were selected more frequently in the online version than in the paper version.

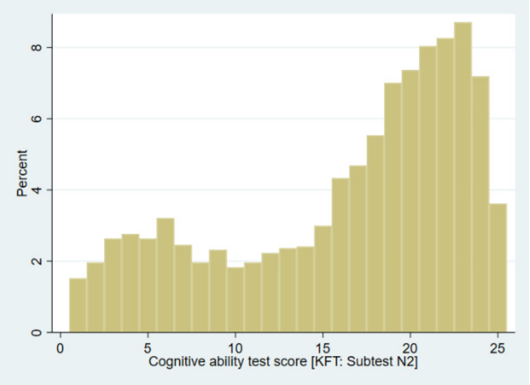
The distribution changes slightly if we exclude all cases who checked multiple answers on one or occasions (Figure 5).

[9] Note that the 8-minute time limit can be mapped to the second in the test's TREE2 web adaptation. We may assume that test administration in paper-and-pencil mode and classroom-setting leads to comparably larger variation in terms of duration of test completion.

There are several (further) potential reasons for reservations as to the comparability of the TREE2 test scores with those of the calibration sample (as reported in Heller & Perleth's test manual of 2000):

- The test among the TREE2 sample was administered in the context of an extensive questionnaire, which may have led to fatigue effects.
- The calibration sample was tested in the year 2000, 16 years prior to the TREE2 sample.
- The original test language, (standard) German, is not the mother tongue of the TREE2 sample.
- In the absence of a device-adaptive test implementation, completing the test on a smartphone rather than a computer may have lead to difficulties in comprehending the test's assignments and navigating through the test.

Figure 5: Distribution of sum scores after exclusion of multiple answers (n= 2598)



The KESS[10] study reports a similar right-skewed distribution (see calculcations by Katja Scharenberg in: Bos et al., 2010). The same is true for own calculations drawing on data of the TIDES[11] study: Although we found similar bimodal distributions for respondents at upper secondary level with basic requirements in Basel-City and Baden-Württemberg (Germany) for the TIDES study, no distribution is as clearly bimodal as that in the TREE2 data. It should be noted, however, that all other results are based on paper-and-pencil versions of the test. Comparability is therefore relatively limited.

Results of cognitive ability tests usually form a (slightly skewed) normal distribution. However, bimodal distributions may be observed when applying the test to different populations. Bimodal distributions may be caused by various factors, such as differences in the sample composition, different educational backgrounds or cultural differences. It is important to investigate such distributions in specific study contexts and discuss the possible causes for the observed distributions.

Table 2: Comparison of the descriptive statistics across different surveys

Survey	TREE2 (Switzerland)	PISA 2000 (Germany)	TIDES (Basel-City, Switzerland)	TIDES (Fribourg, Switzerland)	TIDES (Baden-Württemberg, Germany)
Mode	Web	Paper & pencil	Paper & pencil	Paper & pencil	Paper & pencil
Year	2016	2000	2013	2013	2013
Grade /age	9	15 years	9	9	9
Mean	15.09	14.44	16,77	17.04	15.04
SD	7.10	6.13	5.33	5.07	5.55
Min - Max	1-25	0 - 25	0 - 25	0 - 25	0 - 25
N	3341	6120	1174	856	1258

Sources: PISA 2000 (Kunter et al., 2002); TIDES study, own calculations.

[10] Kompetenzen und Einstellungen von Schülerinnen und Schülern (grade 8).

[11] Transitions In Different Educational Systems, conducted in the Swiss cantons of Basel-City, Fribourg and Baden-Württemberg in Germany. See www.tides-study.ch.

The KESS study reports a similar right-skewed distribution (see calculcations by Katja Scharenberg in: Bos et al., 2010). The same is true for own calculations drawing on data of the TIDES study: Although we found similar bimodal distributions for respondents at upper secondary level with basic requirements in Basel-City and Baden-Württemberg (Germany) for the TIDES study, no distribution is as clearly bimodal as that in the TREE2 data. It should be noted, however, that all other results are based on paper-and-pencil versions of the test. Comparability is therefore relatively limited.

Results of cognitive ability tests usually form a (slightly skewed) normal distribution. However, bimodal distributions may be observed when applying the test to different populations. Bimodal distributions may be caused by various factors, such as differences in the sample composition, different educational backgrounds or cultural differences. It is important to investigate such distributions in specific study contexts and discuss the possible causes for the observed distributions.

Table 2 compares characteristics and distribution of the test scores from different studies. Mean values and standard deviation do not differ significantly, which suggests that the test has worked well.

In view of these findings, a few words of caution are in order here. Even after checking formal validity of the individual test scores and the comparable descriptive parameters, it cannot be ruled out that the CAT performed less well in TREE's online test setting than in earlier paper-and-pencil versions.

In a further step, we therefore checked construct validity to determine the extent to which the test confirms hypothesized correlations. In line with the criterion validity of the CAT manual, our construct validations show the highest (medium to strong) correlations with the grade in mathematics (*tomarkmath*: $r= .22$, $p < .00$) and the score of the extended mathematics assessment[12] administered to the TREE2 sample at baseline (*towlem*: $r= .53$, $p < .00$). In view of the construct's strong reasoning component and of the theoretical concepts that we rely on, this is in line with our expectations. Linguistic constructs yielded rather weak correlations (*tomarklang1*: $r= .10$, $p < .10$ / *tomarklang2*: $r= .07$, $p < .00$ / *toscverb_fs*: $r= .01$, n.s.). Furthermore, medium correlations were found with parental socio-economic status (*tohisei08*: $r= .18$, $p < .00$) and the number of books at home (*tobooks*: $r= .24$, $p <.00$). This supports the validity of the test adaptation, as the CAT is a non-verbal test.

[12] See Nidegger (2019) for details.

6 Conclusion and some words of caution

While the CAT results of TREE2 should be interpreted with some caution, it is encouraging to see that the test scores align closely with those from other studies. Despite the adaptation from paper-and-pencil to web-based test administration and the complex introduction translated into two additional languages, the CAT demonstrated its robustness. To mitigate potential difficulties of comprehension caused by complex test instructions, we consulted with the test publisher to simplify the language and develop an animated explanatory introduction. However, a conclusive assessment of the test's relative performance is difficult without a control group that completed the paper-and-pencil format.

Some disparities were nevertheless observed, particularly among students attending programmes with low academic requirements and those from Italian-speaking Switzerland, suggesting potential areas for improvement in the online testing process. These reservations notwithstanding, we are confident that this innovative test design will provide data users with the tools needed for in-depth analyses of post-compulsory educational pathways that take learners' cognitive abilities at baseline into account.

References

Bos, W., Gröhlich, C., Dudas, D. F., Guill, K., & Scharenberg, K. (2010). *KESS 8 – Skalenhandbuch zur Dokumentation der Erhebungsinstrumente*. Münster: Waxmann.

Brunner, M., Lang, F. R., & Lüdtke, O. (2014). Erfassung der fluiden kognitiven Leisungsfähigkeit über die Lebensspanne im Rahmen der National Educational Panel Study: Expertise (NEPS Working Paper No. 42). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1-22. https://doi.org/10.1037/h0046743

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*(1), 13-23. https://doi.org/10.1016/S0160-2896(97)90011-8

Heller, K., & Perleth, C. (2000). Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision. Göttingen: Beltz Test GmbH.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53-91). The Guilford Press. https://psycnet.apa.org/record/1997-97010-004

Hupka-Brunner, S., Meyer, T., Sacchi, S., Jann, B., Krebs-Oesch, D., Müller, B., . . . Wilhelmi, B. (2023). TREE2 Study Design. Update 2023. Bern: TREE. https://doi.org/10.48350/175367

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., . . . Weiß, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente*. Max-Planck-Institut für Bildungsforschung.

Nidegger, C. (2019). *ÜGK / COFO / VECOF 2016: Competencies of Swiss pupils in mathematics (1.0.0)* [*Dataset*] FORS data service. https://doi.org/10.23662/FORS-DS-1004-1

Scharenberg, K. (2012). *Leistungsheterogenität und Kompetenzentwicklung. Zur Relevanz klassenbezogener Kompositionsmerkmale im Rahmen der KESS-Studie*. Münster: Waxmann.

TREE. (2023). *Transitions from Education to Employment, Cohort 2 (TREE2), panel waves 0-3 (2016-2019) (2.0.0)* [*Dataset*] FORS Data Service. https://doi.org/10.48573/kz0d-8p12

Appendix: Overview of available para-data

In the following, we provide descriptives of the test score's validity status and further para-data that allow us to assess the observed score distributions and the functionality of the test. Variable names and results are drawn from the TREE2 data release as published in June 2023 (TREE, 2023).

cat_status: distinguishes valid from non-valid tests.

```
cat_status — Status of cognitive ability test
----------------------------------------------------------------------------------------------
                                                    |    Freq.     Percent      Valid      Cum.
----------------------------------------------------------------------------------------------
Valid   0 Valid test                                |     3341       87.35      87.35     87.35
        1 Technical problem                         |       10        0.26       0.26     87.61
        2 Incomplete test (break-off, no answer)    |      266        6.95       6.95     94.56
        3 Did not understand the test instructions  |      204        5.33       5.33     99.90
        4 Random/non-serious response pattern       |        4        0.10       0.10    100.00
        Total                                       |     3825      100.00     100.00
----------------------------------------------------------------------------------------------
```

A series of para-data and further auxiliary variables serve to identify the validity of the test scores:

cat_rpattern: identification of non-serious response patterns (e.g., vertical or horizontal straightlining).

```
cat_rpattern — Response Pattern
----------------------------------------------------------------------------------------------
                                                  |    Freq.     Percent      Valid       Cum.
----------------------------------------------------------------------------------------------
Valid   0 No conspicuous response pattern found   |     3819       99.84      99.84      99.84
        1 Random/Non-serious Response Pattern     |        6        0.16       0.16     100.00
        Total                                     |     3825      100.00     100.00
----------------------------------------------------------------------------------------------
```

cat_timeisup: informs on whether respondents had reached the "time is up" screen that was displayed at the end of the test's 8-minute time limit. Serves to identify dropouts/breakoffs.

```
cat_timeisup — Time-is-up page shown
--------------------------------------------------------------
               |    Freq.     Percent      Valid       Cum.
--------------------------------------------------------------
Valid   0 no   |     3298       86.22      87.95      87.95
        1 yes  |      452       11.82      12.05     100.00
        Total  |     3750       98.04     100.00
Missing .      |       75        1.96
Total          |     3825      100.00
--------------------------------------------------------------
```

cat_lastpage: This marker indicates three categories that inform on how far respondents progressed with the test. This variable also serves to identify dropouts/breakoffs.

cat_lastpage — Position at which interview ended

		Freq.	Percent	Valid	Cum.
Valid	1 At intro screen of test	75	1.96	1.96	1.96
	2 Within the test	38	0.99	0.99	2.95
	3 At end of/after the test	3712	97.05	97.05	100.00
	Total	3825	100.00	100.00	

cat_comment: In the comment function, respondents were given the opportunity to mention technical problems, intentional test dropout, or having been unaware of the time constraints. The most frequently expressed comment was that respondents did not understand the test instructions.

cat_comment — Comment by respondent

		Freq.	Percent	Valid	Cum.
Valid	0 Comment not test-related	3685	96.34	96.34	96.34
	1 Did not understand the instructions	115	3.01	3.01	99.35
	2 Was not aware of/had issue with the time limit	7	0.18	0.18	99.53
	3 Had technical problem	10	0.26	0.26	99.79
	4 Cancelled the test	8	0.21	0.21	100.00
	Total	3825	100.00	100.00	