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

TREE Technical Paper Series No. 3

Implementation of a reading speed test in the TREE2 panel survey

Dominique Krebs-Oesch
Stefan Sacchi
Ben Jann

Bern, 2023

Imprint

Published 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, Dominique; Sacchi, Stefan; Jann, Ben (2023). Implementation of a reading speed test in the TREE2 panel survey. TREE Technical Paper Series No. 3. Bern: TREE. Doi: 10.48350/175513

Abstract

This paper focuses on the successful implementation and measurement properties of a bilingual, repeated reading speed test administered to the second TREE cohort (TREE2). Reading speed is a crucial factor for academic success and career trajectories, forming the bedrock for effective learning and information processing. It can also be used as a proxy measure of a person's general reading skills: A high reading speed is a clear indicator of an individual's adeptness at extracting information from texts rapidly and efficiently, while a low reading speed suggests slower reading development.

The inclusion of this test in the TREE2 panel survey hence provides a valuable database allowing for longitudinal investigation of a wide range of interdependencies between the individual development of an important facet of reading literacy, educational pathways and the life course in general.

To this end, TREE has first developed an online version of the originally paper-based reading speed test from NEPS[1] (Zimmermann et al., 2012) which is equally suitable for smartphones, tablets and computers. The digitisation of the pencil-and-paper instrument heralds a new era of test economy and reachability of the target population. In contrast to the original paper-and-pencil test, the innovative adapted version can be applied outside of a traditional classroom setting with a test administrator, thereby greatly simplifying test administration and data collection.

Second, we developed and implemented a French version of the test, allowing us to extend test administration to the French-speaking part of the sample.[2] The French version also allows for comparative cross-language analysis of reading speed. To accomplish this, we relied on a careful multi-stage translation process involving bilingual linguists and proficient translators. All in all, results suggest that we have achieved a high degree of measurement equivalence between both language versions.

The web application also provides valuable process data, such as timestamps measuring response time of individual items or information on early exits from the test. Furthermore, timestamps allow us to identify test dropouts and to check for measurement equivalence across test languages. This provides us with the opportunity to analyse appropriate test administration and the (formal) validity of individual test scores.

The publication of the test data in the 2023 TREE2 data release (TREE, 2023) provides researchers with the opportunity to more comprehensively investigate the role of reading speed in educational and professional contexts and to gain new insights into its impact on individual and societal development. The present paper provides a comprehensive description of the adapted reading speed test, including selected results regarding its reliability, criterion validity and cross-language measurement equivalence.

[1] National Education Panel Survey (Germany).

[2] Unfortunately, funding and resources did not allow for the development of an Italian version of the test. Therefore, the Italian-speaking sample of TREE2 (<10% of the overall sample) did not undergo the test.

Table of contents

Overview 5
1 What is reading speed? 8
2 Implementation of a reading speed test in TREE2 9
2.1 How does the test work? 9
2.2 Implementation of the test in the NEPS 9
2.3 Research interests 10
2.4 Adapting the test to TREE2 purposes 11
2.4.1 Linguistic adaptations 11
2.4.2 Adapting the test to web mode 12
2.5 Randomised test administration from panel wave 3 onwards 15
2.5.1 Available reading speed test sample 18
2.5.2 Test-related variables 22
3 Empirical analyses 22
3.1 Formal validity of test data 22
3.2 Descriptives 24
3.3 Ceiling effects 25
3.4 Test reliability and validity 27
3.4.1 Split-half & retest reliability 27
3.4.2 Criterion validity 27
4 Conclusions and some words of caution 28
5 References 30
Appendices 32
Appendix A: T-test tables for reading speed test time differences 32
Appendix B: Overview of variable names and labels, value labels and annotations across datasets 38
Appendix C: Assignment of missing value codes and sample statistics in Stata and SPSS dataset 41
Appendix D: Weighted descriptives of Table 3 (page 24) 42
Appendix E: Correlations with validation criteria 43
Appendix F: Data structure 44

Overview

Why implement a reading speed test?

Implementing a reading speed test is a substantial "added value" to the TREE2 panel survey for several reasons. The test provides a measurement of a fundamental aspect of reading literacy. On the one hand, it complements other skills measures (in mathematics and cognitive skills) that do not cover reading skills. On the other hand, a measurement of reading skills is also useful in view of TREE's multi-cohort design and its potential for cohort-comparative analyses of post-compulsory trajectories and transitions. The focus of skills measurements at baseline in the first TREE cohort (TREE1) was on reading literacy (BFS & EDK, 2002), whereas in TREE2 it was on mathematics skills (AES[3] 2016). Although the reading speed test implemented in TREE2 can by no means replace comprehensive measures such as PISA[4] (Angelone & Keller, 2019), it nevertheless provides TREE data users with an instrument to control for at least basic reading skills across cohorts, thus improving the analytic value of cohort-comparative research.

Why is the NEPS reading speed test suitable for TREE2?

When we looked for a test suitable for TREE2, design requirements were such that the instrument had to be short, validated, web-based and self-administered (CAWI/CASI). Furthermore, it should also be suitable for repeated measurement in later stages of the cohort's trajectory. Against the background of these requirements, we considered the NEPS[5] reading speed test (Zimmermann et al., 2012) to be particularly suitable for the implementation in a web application, since the assessment of individual sentences as being correct or incorrect is easy to technically implement and to instruct. As to the requirement of suitability for repeated measurement, the test has certain limitations with respect to ceiling effects. According to results from NEPS longitudinal use of the test among different age cohorts, ceiling effects occurred with increasing age of respondents (e.g., 5.4% of the adults completed all items correctly; Zimmermann et al., 2014). However, as the test differentiates well in the lower range of reading ability at any age and is suitable for different age groups, we may assume that it is appropriate for longitudinal use in the TREE2 panel. One of the major advantages of the test is its length of only two minutes, which minimises additional survey burden for panel respondents.

3 See Konsortium ÜGK (2019)

4 The results of the two tests are not fully comparable for two reasons. On the one hand, PISA is a comprehensive measure of reading literacy, which cannot be fully captured by a short reading speed test. On the other hand, there are slight differences as to the point in time of measure: The Swiss PISA 2000 sample (i.e., the baseline sample of the TREE1 cohort) has been tested in the last year of lower secondary education (i.e., shortly prior to the end of compulsory school), whereas the reading speed test in TREE2 has been administered one year after the cohort has left compulsory school.

5 National Education Panel Survey (Germany).

What does the test measure?

The test primarily measures decoding speed, which comprises two basic components of reading, namely reading speed and decoding accuracy. Both reading speed (number of words or sentences read within 2 minutes) and decoding accuracy (i.e., proportion of correctly comprehended words respectively correctly assessed items) are part of the automatised text processing while reading and indispensable for text comprehension. The test itself therefore consists of a list of simple statements that are to be read as quickly as possible and to be assessed as true or false.

How has TREE adapted the test?

The test employed by TREE is basically a literal substitution of the NEPS paper-and-pencil administration by a web-based version that we have administered in panel wave 1 and from panel wave 3 onward. We implemented two device-specific adaptations. For desktop and laptop devices, the test's 51 items are displayed on five screens linked with "next" buttons, each of which can be completed without scrolling. In the smartphone/tablet version, all items are displayed on one and the same screen that can be thumb-scrolled. For the German-speaking part of the sample we adopted the German NEPS version of the test, restricting adjustments of content to a few minor "helvetisms".[6] In order to be able to administer the test to a maximum share of the TREE2 sample, we translated the test to French.[7]

What should be kept in mind when working with the test data?

The longitudinal implementation of a reading speed test in TREE2 provides researchers with a wide range of opportunities for analyses. This paper describes a number of them (see 2.3 Research interests), some of which also open up thanks to a partly randomised test administration.

One straightforward approach to use the test data is to rely on the cross-sectional measurements available for the full sample in panel wave 1 and for a split-half sample in panel wave 3. These data can be used for cross-sectional analyses or to analyse intra-individual changes in reading speed (based on one randomised split-half sample) between age 17 and 19.[8]

Depending on the research objectives, the relevant data are to be found in one of the following datasets:

- For wave-specific analyses relating to panel waves 1 and 3 and to measures of intra-individual stability of reading speed, the test data are included in the respective general wave-specific data files ('*TREE2_Data_Wave_1_v2 and TREE2_Data_Wave_3_v2*').

[6] Idiosyncratic terms in the Swiss national languages that are only in use in Switzerland.

[7] For reasons of research economy, we refrained from translating the test to Italian, TREE's third survey language. This concerns less than ten percent of the sample.

[8] The mentioned split-half sample is the predefined wave split-half sample. For further information on the test administration design and the two split-half samples see Section 2.5.

- For analyses pertaining to the role that reading speed plays for the attainment of an upper-secondary diploma and the transition from upper-secondary education to post-secondary education or the labour market, we recommend drawing on the data file '*TREE2_Data_rs_graduationyear_v2*'.[9] This file comprises test data that have been collected shortly prior to respondents' expected graduation from formal upper-secondary education.[10]
- The dataset '*TREE2_Data_rs_v2*' includes all cases that comprise valid data from the reading speed test. The file also includes test-related para-data and the data pertaining to the single test items. By means of the complementary Stata syntax '*Stata-Syntax TREE2_Do-file_rs_v2*', scholars working with the data may check how test scores were generated and validated.
- For a detailed documentation of the partially randomised design pertaining to the longitudinal test administration we refer to Section 2.5.
- A detailed overview of the test-related variables can be found in Appendix B, while more information on the structure of the datasets can be found in Appendix F.
- It should be noted that the test results vary in terms of sample selectivity. Due to non-response and randomized test administration, some results are available for specific subsamples only. Unbiased population estimations therefore require the use of suitable survey weights and/or data imputations (see Section 2.5.1 for details).

[9] See folder 1-4 'Reading Speed Test' in the TREE2 2023 data release (TREE, 2023)

[10] At the time of data release in 2023, however, only a portion (about half) of those tested in the final year according to our test design (see Section 2.5) will have been tested. For analyses on the transition from lower- to upper-secondary education, further waves must be awaited until a large proportion of young adults have entered the world of work or imputation should be carried out.

1 What is reading speed?

Reading speed is an indicator of a basic (automated) reading process in which the ability to decode words accurately (accuracy) and speed through automation (automaticity) interact. In the first place the automation of the hierarchical decoding processes liberates cognitive resources for advanced processes of comprehension (Rosebrock & Nix, 2006). According to Sturm (Sturm, 2011), reading speed results from the interplay of accurate and automated decoding. Reading speed as a basic and cognitive element of reading skills is relevant for the entire reading process. It is significantly related to efficiency in accessing the semantic lexicon and essential for the quality of the reading process (Perfetti, 1985; Perfetti & Hogaboam, 1975, in Hartmann 2012). Slow readers who are unable to retain (detailed) information in their working memory therefore encounter comprehension problems (Pissarek, 2018; Rosebrock & Nix, 2006). Accordingly, a low level of automation of the decoding processes leads to poorer text comprehension and hampers the entire process of understanding (Rosebrock & Nix, 2006, in Autorenteam Kompetenzsäule, 2020). This is not only the case with young inexperienced readers. Inter-individual differences in reading speed were also found in older age-groups (Autorenteam Kompetenzsäule, 2020). It is therefore not surprising that the basic processes of reading predict differences in reading comprehension (Artelt et al., 2001; Hahnel et al., 2017). The extent to which there is a causal link between reading speed and reading comprehension has not yet been fully clarified. However, it is consensually assumed that there is a reciprocal relationship, as evidence shows that good readers read faster than poor readers (Rosebrock & Nix, 2006).

Furthermore, a high correlation of reading speed with cognitive skills is assumed, as slow readers lack the necessary processing capacity for the formation of mental representations of the read text. This hampers the formation of coherence at the level of the text (Artelt et al., 2001, pp. 50,; cited in Hartmann, 2012). Therefore, a low reading speed is presumably the result of too much cognitive effort in word recognition and simultaneous comprehension problems at a higher processing level (Hartmann, 2012).

This condensed overview of the relevance of reading speed for reading literacy as a whole is of course by no means exhaustive. Nevertheless, it underlines its importance when it comes to analysing developmental trajectories such as those investigated by the TREE study.

2 Implementation of a reading speed test in TREE2

2.1 How does the test work?

The reading speed test provided by NEPS is based on the principle of the Salzburger Lesescreening (SLS, see Auer et al., 2005) and consists of 51 individual sentences. The sentences are short, simply structured and must be classified by the respondents as being correct or incorrect in terms of content. The sentences are ordered in ascending length and start with five words, while the longest sentences measure 18 words. The test takes two minutes, and the number of correctly solved test items (sum score) is counted as an indicator of reading speed. Since the aim of the test is to capture automated reading processes, the sentences reflect everyday knowledge and do not contain a "knowledge component" (for further information on the NEPS test construction, see Artelt et al., 2001; Zimmermann et al., 2012).

Figure 1: Exemplary test items[11]

	richtig	falsch
In einer Garage findet man immer eine Badewanne.	☐	☐
Die Person, die bei einem Fußballspiel auf die Einhaltung der Regeln achtet, nennt man Schiedsrichter.	☐	☐

As mentioned above, the reading speed test co-measures information processing in reading by means of asking respondents to verify or falsify the sentence. Since the test primarily differentiates individual differences in the lower performance range, it can be used for different age groups. As the time limit of two minutes is the same for all target groups, the performance is also comparable across different age groups. However, the older the tested individuals are, the more likely ceiling effects will occur (Zimmermann et al., 2012). We implemented the test despite this shortcoming, as our research interest lies in the monitoring of longitudinal changes in reading skills among individuals with restricted reading performance (see Section 3.3 for more detail on ceiling effects).

2.2 Implementation of the test in the NEPS

In the NEPS[12] the sum score of the reading speed test is primarily employed as a reading-related control measure. It allows to analyse developmental trajectories, as domain-specific (reading) skills can be measured independently of initial differences in decoding speed. For this reason, it is advisable to first administer the test in the earliest possible panel wave. Depending on the panel design, the test can then be repeated at looser intervals on the occasion of later panel waves.

[11] "In a garage, one will always find a bathtub."
"The person who makes sure that the rules are obeyed in a football match is called a referee."
(correct vs. false). See Autorenteam Kompetenzsäule (2020).

[12] National Education Panel Survey (Germany).

The NEPS is also interested in the longitudinal measurement of reading speed. Due to the ceiling effects of the test in the older age groups, NEPS has worked on an extension of the test, which was first administered as a repeated measurement in 2019.[13] Furthermore, NEPS also switched test modes from paper-and-pencil to computer-based administration. In order to avoid ceiling effects, an extended computer version with 72 (instead of 51) items is available for repeated measurement as of 2020 (Autorenteam Kompetenzsäule, 2020). TREE refrained from adopting the extended NEPS version for two reasons. On the one hand, the extension was not (yet) available when needed; on the other hand, TREE prioritized comparability across points of measurement and therefore continued to administer the original test with 51 items.

The NEPS test reports medium-size correlations between reading comprehension and reading speed in 5th and 9th grades ($r_{5grade} = .34$ / $r_{9grade} = -.60$). Furthermore, reading speed in 5th grade is only moderately correlated with reading speed four years later in 9th grade ($r = .46$).[14] Similar results are reported with regard to tests administered among adults. The correlation between reading speed and the reading literacy test administered by PIAAC is at $r = .55$.[15] Overall, correlations between the NEPS reading speed test and reading comprehension remain at medium-size after 9th grade ($r \approx .55$). Reading speed therefore allows, at best, for only very approximative control of reading skills such as they were measured among the first TREE cohort (TREE1).

2.3 Research interests

The test design developed by TREE aims at covering a maximum range of analytical interests, the three most important being:

1. *Measuring reading speed development over time*
   Repeated cross-sectional measurements are used to analyse the intra-individual change or stability of reading speed over time. These measurements take place in the first and third year after the end of compulsory schooling for a part of the sample.

2. *Measuring reading speed prior to upper-secondary graduation*
   For another part of the sample, the test is administered to respondents in the year prior to their graduation from upper-secondary education, irrespective of the panel wave at which this is the case. Depending on the duration of the programme respondents are enrolled in, the time span between the initial measurement (in panel wave 1) and the measurement prior to graduation may cover several years. The measurement prior to graduation allows for a large range of analyses with respect to the transitions from upper-secondary to tertiary level education or to the labour market.

[13] Personal communication from Karin Gehrer, member of the NEPS test development crew.

[14] Personal communication from Karin Gehrer, member of the NEPS test development crew.

[15] Own calculations. PIAAC = Programme for the International Assessment of Adult Competencies.

3. *Measuring reading speed close to the baseline survey (in analogy to TREE1)*
   Contrary to TREE1, the TREE2 baseline survey included testing of mathematics, but not reading skills (see Hupka-Brunner et al., 2023 for more detail). In order to increase cohort comparability, the first measurement of reading speed in TREE2 was implemented at the earliest moment possible (in panel wave 1).

2.4 Adapting the test to TREE2 purposes

The original test provided by NEPS in 2017 was designed for paper-and-pencil mode and mostly for administration in a school-based setting. TREE therefore developed two major adaptations of the original test. On the one hand, we adapted the German test to the Swiss-German context and developed a French version, in order to be able to test both the (Swiss-)German and the French-speaking part of the TREE2 sample.[16] On the other hand, we converted the test to web mode.

2.4.1 Linguistic adaptations

Adaptations in German

Although German is taught in all Swiss-German schools, colloquial helvetisms[17] are widespread in German-speaking Switzerland. In consultation with the NEPS, minor adjustments to the original German items were therefore made to avoid disturbances of the respondents' reading flow, e.g., by accounting for helvetisms and national designations.[18]

Translation to French

There is hardly any scientific literature on translations of reading speed tests which not only measure speed, but also basic reading comprehension (classifying sentences to be correct or incorrect). The IReST[19] project provides some guidance. In this project, (bilingual) linguists drafted texts in different languages, comparing them in terms of content, length, difficulty and linguistic complexity. In doing so, the IReST linguists drew up texts of almost identical length (+/-2 characters) in no less than 17 languages (see also Trauzettel-Klosinski & Dietz, 2012). In line with their work, TREE primarily aimed to translate the original test items as literally as possible, mandating two translation teams to this end. In a second step and drawing on the services of a bilingual (German-French) linguist, we validated and adapted the items with regard to

- the number of characters (including visual length of text in both languages),
- the location of keywords for text comprehension and

[16] Unfortunately, funding and resources did not allow for the development of an Italian version of the test. Therefore, the Italian-speaking sample of TREE2 (<10% of the overall sample) did not undergo the test.

[17] Idiosyncratic terms in the Swiss national languages that are only in use in Switzerland.

[18] E.g., «Fahrprüfung» instead of «Führerscheinprüfung» (driver's license) or «Bundesrat» (Federal council) instead of «Bundeskanzler» (Federal Chancellor).

[19] International Reading Speed Texts. See https://www.vision-research.eu/index.php?id=641 for more detail.

- the linguistic complexity of the sentences.

The linguist's expertise was particularly valuable with respect to adequately reflecting the everyday language of French-speaking Switzerland. This is crucial when it comes to avoiding disruption of the automated decoding flow owing to the use of uncommon wording or phrasing.

Both in the pretest phase and in the main field of panel wave 1, the response time of each item was compared using time stamps[20] in order to evaluate the translation. Apart from a few exceptions, the evaluation showed that item-specific response time varied little across languages.

A final evaluation after panel wave 3 confirmed these results. Three appended tables display item-specific response time in seconds for both languages (see Appendix A). We draw on T-tests to examine the extent to which response time differs across the two languages. The first table (Table A1) displays response time data that are pooled for panel waves 1 and 3. Owing to the larger data base (especially for the last items), this table is discussed in more detail below. For the sake of completeness we also list the panel wave-specific response times in two additional appendix tables (A2 and A3). As the effect sizes hardly differ from the pooled model, the latter are not commented on below.

Due to the large number of cases and the associated frequent significance of the tests, Cohen's *d* is useful to classify even small differences in response time. Table A1 shows only marginal effects for the most part of the items (Cohen's $d < .2$). However, 19 items show small differences (Cohen's $d >= .2$). Only one test item (item 46) displays a medium effect size (Cohen's $d = .56$).[21] The sign of Cohen's *d* indicates in which language response time was longer: positive values reflect a higher mean score in the first group (German), while negative values reflect a higher mean score in the second group (French). Among the 20 items with small to medium effect size, exactly half have a negative and half a positive Cohen's *d* value. This can be interpreted as a further indication that the translation worked out well. We take the overall small to negligible effect sizes and the linguistic distribution of the longer sentences as an indication that the French version is valid and the French reading speed test is equivalent to the original German version.

2.4.2 Adapting the test to web mode

We developed the web adaptation of the paper-and-pencil instrument in close cooperation with NEPS. The paper-and-pencil version allows respondents to see all the items on a double page and to tick "right" or "wrong" by hand. For reasons of screen size, this layout cannot be replicated in the web version.

[20] Time stamps were set for the assessment (correct or incorrect) of each item. In doing so, response time for each item can be compared across languages.

[21] How to interpret Cohen's *d* effect size: .2 = small, .5 = medium, .8 = large (Cohen, 1988).

TREE's web adaptation accounts for the device with which the test is completed. If the test is completed on a desktop or laptop computer, we split the test across five consecutive screens with an equal set of items that do not require scrolling. Clicking on the "next" button navigates from one screen to the next. We are thus confronted with the disadvantage that respondents do not see the entire test at one glance as they would in the paper version. However, by maximizing the number of items per screen (und thus minimizing the number of necessary screens for the entire test) we aim at counteracting scrolling heterogeneity[22], which may have an effect on the overall timing. The adaptation for smartphones and tablets displays all 51 items on a single page.[23] For these devices, scrolling is unavoidable because of screen sizes, but more homogeneous and common (using fingers). We therefore refrained from using the layout for desktop/laptop devices, as a combination of scrolling and the "next" button on smartphones or tablets is rather cumbersome and costs further test time.

The test setting of the web version differs substantially from the standard setting of the paper-and-pencil mode (self-administered vs. instructed). The test instructions therefore had to be thoroughly revised in order to achieve a functional equivalent to the classroom setting in which the paper-and-pencil version is usually administered. In the latter case, test administrators are present, explain the test, answer students' questions and control the test time by means of a stopwatch. As a functional equivalent to the stopwatch, TREE's web adaptation implemented an integrated automatic time limit, which communicates, by means of a pop-up screen at the end of the two minutes test time, that test time is up and the test is over.

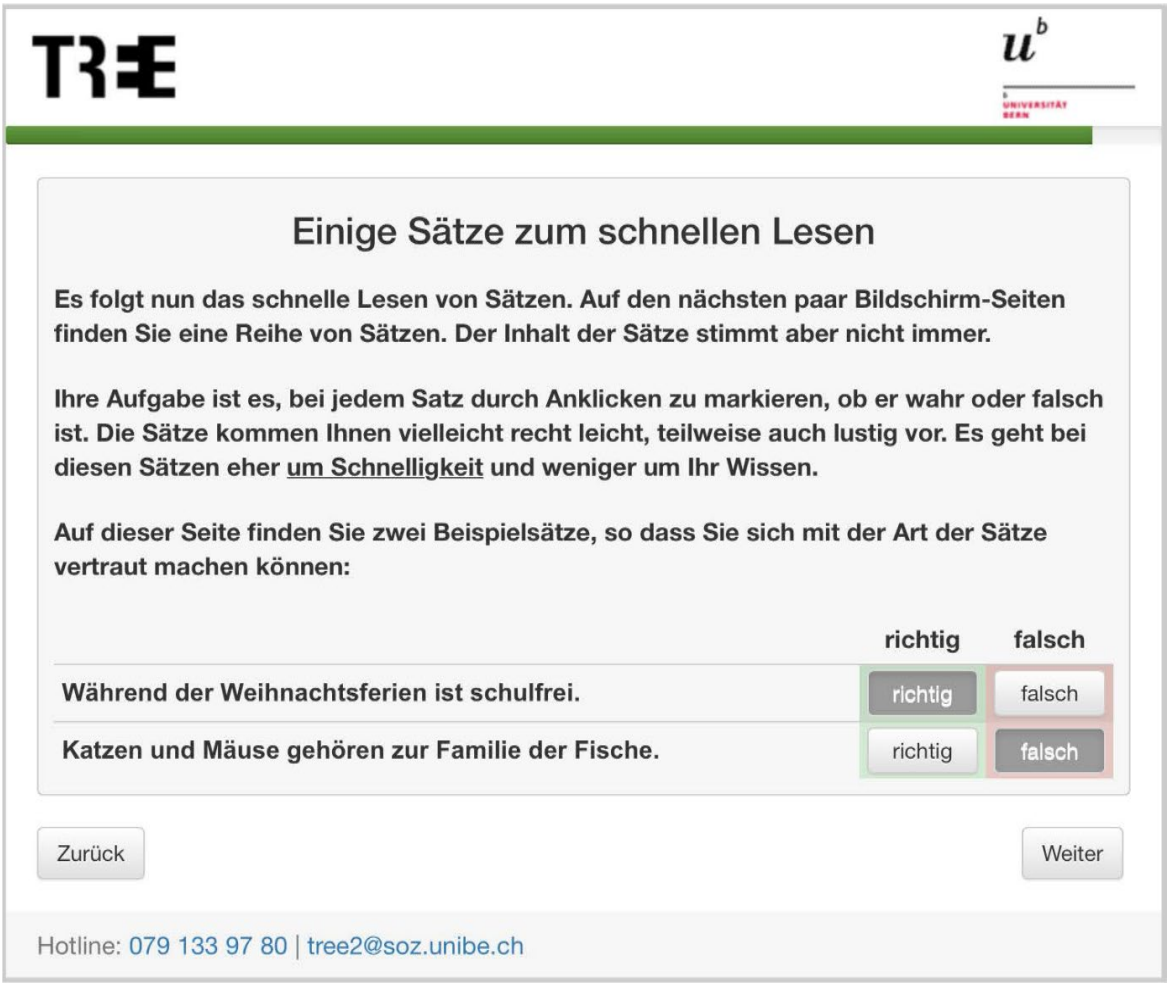
Instructions on the required speed and the time limit of two minutes were particularly crucial, as the TREE survey does not comprise any other questions that impose time limits. While the title on the first instruction page, "A few sentences for quick reading" (see Figure 2), does refer to the speed aspect, it attempts to avoid the explicit mention of a "test" and the two-minute time limit. The latter is mentioned on the second instruction screen only (see Figure 2). We thus attempted to avoid respondents' premature termination of the test.

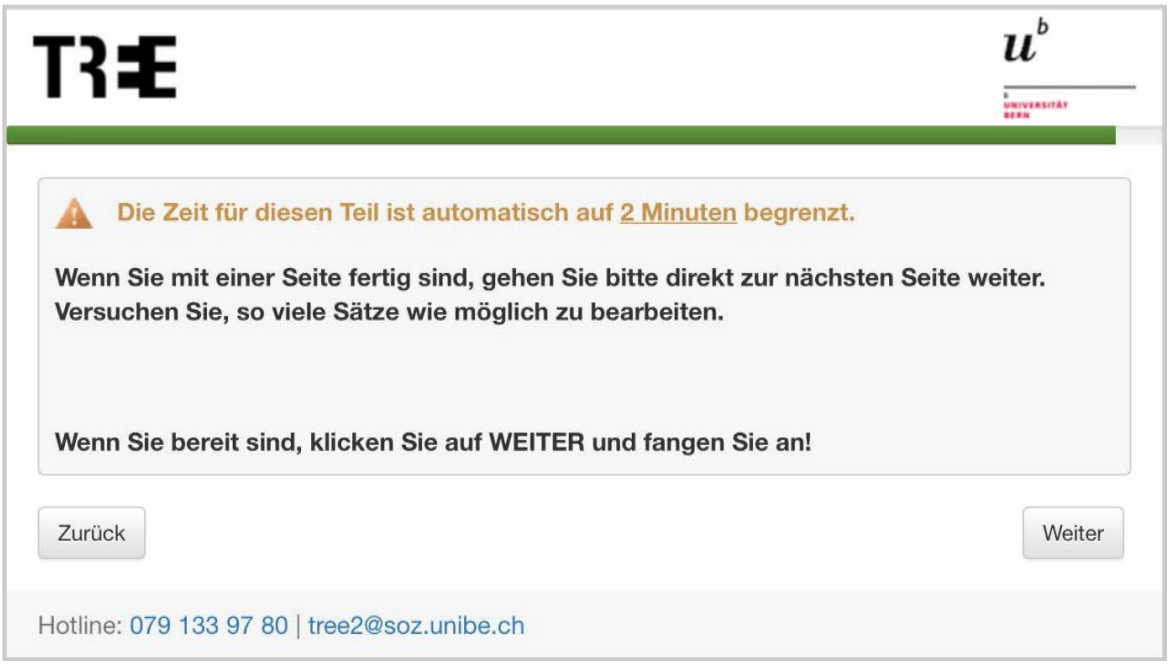
A further important aspect is the provision of a direct contact (hotline number and email address in the footer of the screen), in the event that respondents have questions or encounter difficulties while completing the test. However, this offer was not used.

[22] E.g., scrolling by means of the cog of the mouse vs. vertical scrolling bar vs. scrolling by touchpad.

[23] When first implementing the test in panel wave 1 (2017), we initially planned to restrict the test to desktop/laptop computers. However, it quickly became apparent that a large part of the respondents (≈ 50% in panel wave 1 and ≈70% in panel wave 3) completed the questionnaire on the smartphone.

Figure 2: Introduction to the web-based test (screenshots of the German version's first and second screen)

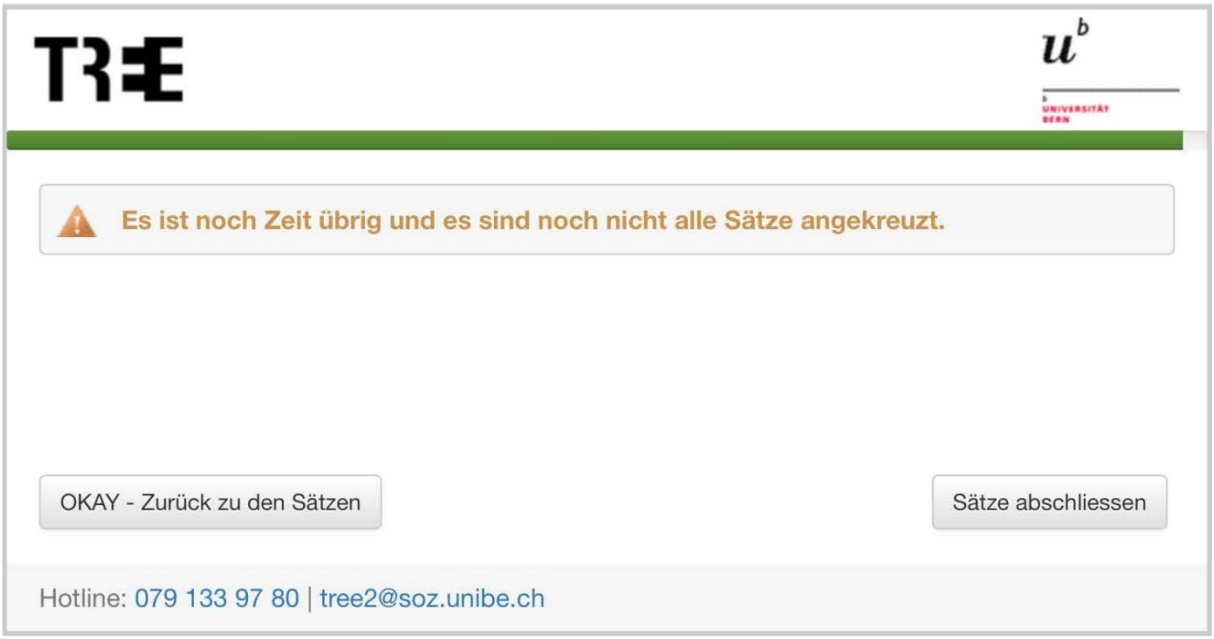




In order to further maximise comparability with the paper version, we also

- implemented large "right/wrong" buttons for easy clicking;
- kept the line breaks identical to the paper version;
- gave respondents the opportunity to complete hitherto uncompleted items, provided the two-minute time limit had not yet been reached (Figure 3). If respondents then clicked back into the test, they were automatically shown the screen with the first uncompleted item. This is important because the paper-and-pencil version always shows the whole test, but the screens only show part of the test items.

Figure 3: Reminder of uncompleted items



2.5 Randomised test administration from panel wave 3 onwards

The reading speed test was first administered in 2017 in the first panel wave of TREE2. The aim was to obtain a baseline measurement of reading speed for the entire TREE2 sample. Subsequently, the test was randomly administered either in the third panel wave (2019) or in the year before the expected completion of post-compulsory education. This allows us to meet all analytical goals outlined in Section 2.3, while at the same time minimizing the survey burden through repeated test administration. In this Section, we describe the test administration design that we applied in more detail.

The randomised administration of the test is illustrated in Figure 4: First, the initial sample available for panel wave 3 was randomly divided into two split-half samples. The cloud of dots in Figure 4 represents the TREE2 sample, which remains unchanged across all waves (i.e., non-response and attrition are neglected for the purpose of illustration; see Section 2.5.1). In one of the two equally sized sample splits (lower half in Figure 4), the reading test was then administered unconditionally in panel wave 3 (i.e., "pre-defined wave split-half"). As of panel wave 3

(and onward), respondents in the other split-half sample are administered the test only on condition that they attend the final year of the upper-secondary education programme they are enrolled in (see blue dots outlined in red in the "graduation year split-half"; upper half of Figure 4).[24] The adopted administration design implies that from panel wave 4 onward, respondents in the "unconditional" split-half sample (lower half of Figure 4) will not undergo the reading speed test even if they satisfy the condition of being in their last year of upper-secondary education.[25] In wave 3, only respondents who were in the graduation year sample split, but not in their final programme year in this wave do not receive a test (grey dots in column "3"). On the other hand, the design also implies that from panel wave 4 onwards, there are respondents who are in their final year but are not administered the test (grey dots outlined in red in the pre-defined wave sample split; lower half of columns "4" to "6").

It should be noted that some respondents had been already in their final year of training in panel wave 2 (*EBA/AFP* VET programmes).[26] However, the longitudinal test administration design was only implemented in panel wave 3, which is why the respective cases (138 in panel wave 2) are marked as "not administered" in both sample splits of Figure 4. We were able to identify these cases retrospectively on grounds of training information they reported.[27] It should be noted that a considerable number of the cases in question enrol in 3-4-year VET programmes after EBA/AFP graduation and thus will be tested at a later date according to the test sampling design. Depending on the analysis that data users may want to conduct, this should be taken into account by controlling for the programme attended in panel wave 2.

The test administration design outlined so far ensures that all respondents in the panel take the test exactly twice and with a minimum test interval of 2 years.[28] On the one hand, this is optimal with respect to minimising survey burden and potential retest effects (Scharfen, Peters, et al., 2018). On the other hand, the design covers all three analytical interests listed in Section 2.5:

[24] Respondents provide this information during the CATI interview preceding the complementary questionnaire survey (base questionnaire, see Hupka-Brunner et al., 2023 for more detail).

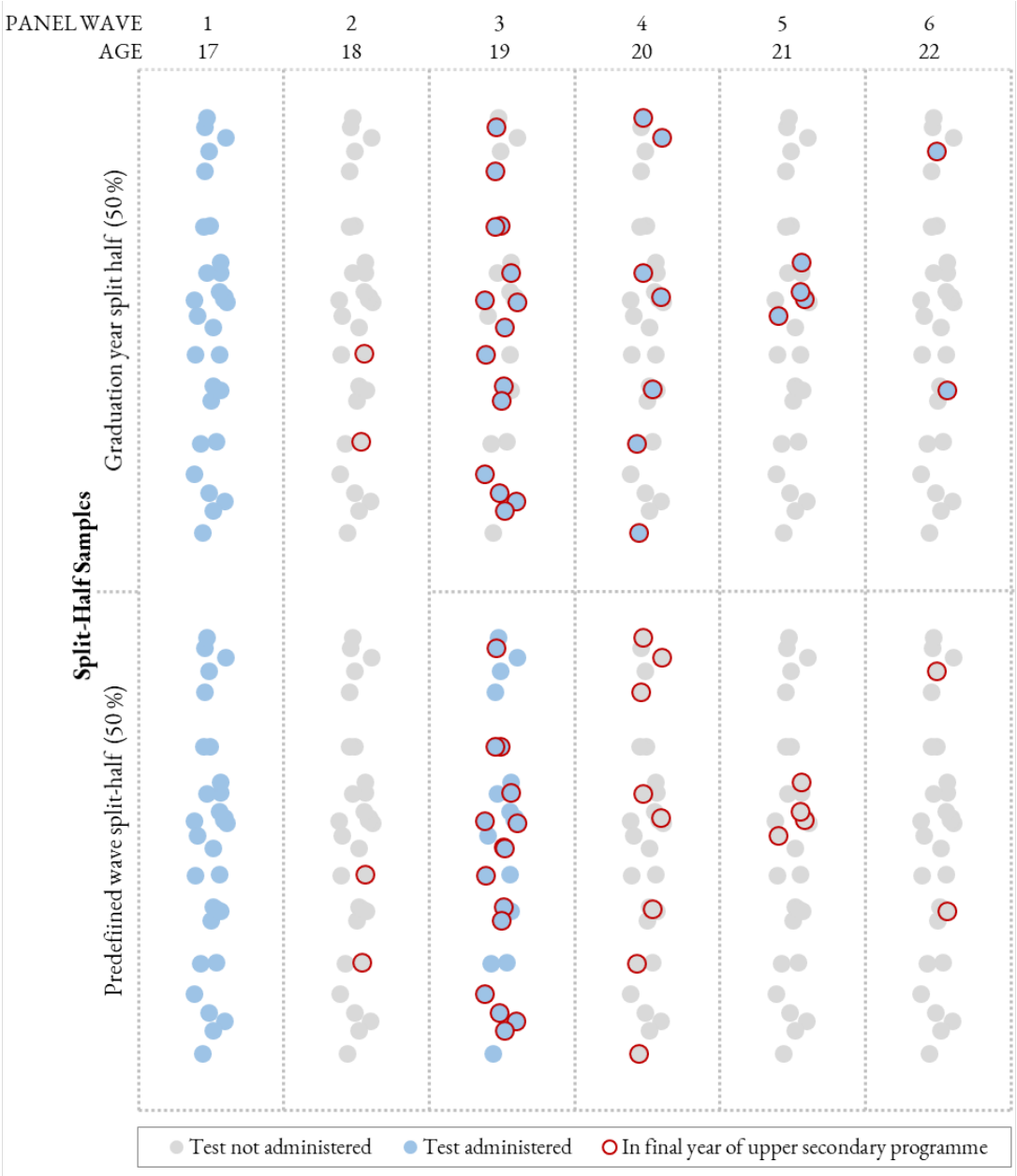
[25] As of panel wave 3, the test is hence administered to respondents as soon as they report to be (currently) in their last year of upper-secondary programme. This may be the case more than once in case of dropout/change of programme.

[26] Two-year VET programmes leading to a *Eidgenössischer Berufsattest (EBA)/Attestation fédérale professionnelle (AFP)*.

[27] For details on how to identify these cases in the datasets, see annotations of the variables *rs_assigned* and *rs_graduationyear* in Appendix B.

[28] Apart from the length of the test interval, substantial retest effects can also be largely excluded due to its simplicity (Scharfen, Blum, et al., 2018).

Figure 4: Partially randomised test administration design



ad 1) It comprises a split-half sample with two measurements in pre-defined panel waves, administered at a fixed, two-year test interval (panel waves 1 and 3, i.e., at the age of approximately 17 and 19 years). This allows for analyses of intra-individual change or stability of basic reading skills by taking into account respondents' educational trajectories.

ad 2) For the "graduation year" split-half sample, the design provides a measurement of reading speed — a basic form of reading literacy — shortly before the end of upper-secondary education. It thus captures a potentially relevant predictor for modelling successful graduation from this level of education and for the transition processes to further stages of the educational and/or occupational career.

ad 3) In view of analyses comparing the two TREE cohorts, the measure available for the entire sample at panel wave 1 provides a proxy control of reading skills shortly after compulsory school.

The test administration design implies that we dispose of one split-half sample each with approximately half of the realised tests for each of the analytical purposes mentioned under ad 1) and 2) above. Given that the test administration is randomised , however, it may be a promising strategy to impute the missing test scores (see e.g., Van Buuren, 2018), so that the whole sample may be used. On the one hand, we may impute missing test scores for wave 3 in the "graduation year" sample split in order to obtain complete measurements for the full sample of panel wave 3 (grey circles in the upper part of wave 3 in Figure 4). On the other hand, the missing graduation year test scores in the "pre-defined wave" sample split may also be imputed (grey circles outlined in red on the lower part in Figure 4).[29] Note, however, that with the data available in the 2023 release, only the former imputation is feasible.

With regard to imputation, we may exploit the available unconditional measurement from panel wave 1. To this end, it is also favourable that the share of participants who were in their last year of training reaches its maximum in panel wave 3 (43%; n=2,551). Hence, the proportion of test data to be imputed is significantly lower than it would have been if the test had been conducted in another panel wave.

2.5.1 Available reading speed test sample

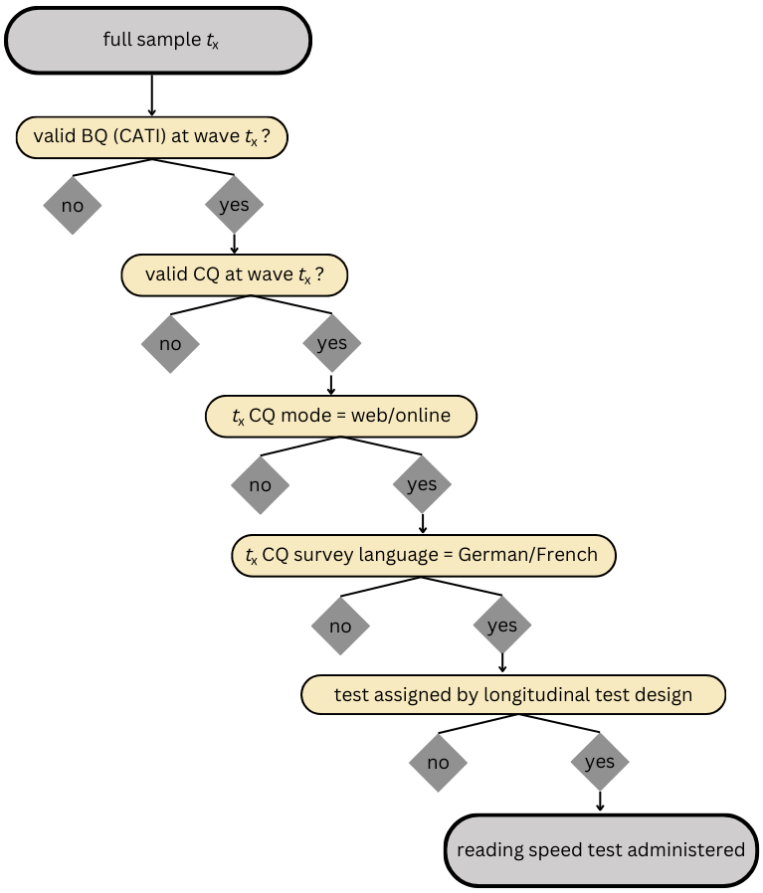
In the following, we provide an overview of the frequency of various reasons for non-administration or non-response to the reading speed tests. The full initial sample consists of school leavers who have completed the TREE2 base questionnaire (BQ) of the respective panel wave.[30] This is a necessary precondition for participation in the reading speed test. For these samples, suitable survey weights are available in the TREE2 2023 data release (TREE, 2023), providing an approximate correction of sample bias in any population estimates (see Sacchi, 2023 for details). The administration of the reading speed test was further restricted to respondents who completed the BQ via CATI (thereby excluding a limited share of respondents to the BQ's paper-and-pencil version) and subsequently also completed the online version of the complementary questionnaire (thereby excluding both non-respondents and respondents to the CQ's paper-and-pencil version). Furthermore, the administration of the test is restricted to the German or French language version of the CQ. The reading speed test is then administered to the remain-

[29] The randomised test administration implies that the MAR assumption, which is critical for imputations, is fulfilled, provided that the year of training (in the case of imputation of missing test scores among the "pre-defined wave" split-half of the panel wave 3 sample) or the survey wave corresponding to the graduation year (imputation of missing test scores for the last programme year) is controlled. The variables required for this from the reading test dataset are *(t3)rs_graduationyear* (flag: in graduation year) and *(t3)rs_split* (identifies the split-half samples). Note, however, that there are also other sources of missing test scores (i.e., non-response to the complementary questionnaire), for which the MAR assumption may be more critical (see Table 2 in Section 2.5.1).

[30] For details on the TREE2 survey design and its questionnaire parts, see Hupka-Brunner et al. (Hupka-Brunner et al., 2023).

ing respondents according to the longitudinal administration design described above (see Section 2.5). A flow chart of the individual selection steps ultimately leading to the test administration is displayed in Figure 5 (for the respective wave-specific frequencies of cases, see the tables below).

Figure 5: Sampling design of the TREE2 reading speed test



It should be noted that there are missing test data even among the respondents to which the test was administered. On the one hand, this was due to occasional technical problems with the test.[31] On the other hand, some respondents dropped out or refused to take the test altogether. Table 1 provides overview of the reading test sample available for panel wave 1.

[31] On the one hand, technical problems were identified by means of para-data, for example when the test was not completed and no timestamps were available for the test items. On the other hand, some respondents reported in their final comments that instead of a test, only empty browser pages were displayed.

Table 1: Test response parameters in panel wave 1 (2017)

	Survey language [1)]			Total sample	Non-response and exclusions
	German *n*	French *n*	Italian *n*	*n*	*(%)*
Initial sample with valid base questionnaire (BQ)	5533	1995	443	7971	
of which paper-and-pencil mode	– 288	– 94	– 20	– 402	–5.0 %
Sample with valid BQ data	5245	1901	423	7569	
of which incomplete CQ	– 1603	– 493	– 112	– 2208	–29.2 %
of which valid paper-and-pencil mode of CQ	– 691	– 328	– 62	– 1081	–14.3 %
Sample with complete CQ data	2949	1081	250	4280	
of which test not administered due to language[2)]	0	0	– 250	– 250	–5.8 %
Initial reading speed test sample	2949	1081	0	4030	
of which technical problems with test administration	– 51	– 25	/	– 76	–1.9 %
of which test refused, not valid or missing	– 106	– 34	/	– 140	–3.5 %
Realised sample with valid test data	2792	1022	/	3814	

1) Respondents are at liberty to choose one of the three survey languages. With a few exceptions the survey language corresponds to the majority language of the language region. In rare cases, the survey languages of BQ and CQ differ. 2) No test is available for Italian.

Test non-response is mainly due to non- or incomplete response to the complementary questionnaire (at the end of which the test is administered) and to the relatively frequent administration of paper-and-pencil modes of the questionnaires. To a lesser degree, the exclusion of the Italian language area and test-related refusals or technical problems contribute to non-response. The realised sample of 2792 cases corresponds to the blue dots in the leftmost column (panel wave 1) of Figure 4.

In approximately equal parts, test non-response in panel wave 3 is due to CQ non- or incomplete response and to non-administration of the test based on the test sampling design. The exclusion of the Italian language area and questionnaires administered in paper-and-pencil mode are of less importance, the latter even in comparison to panel wave 1.

It should be noted that the survey weights available in TREE2 are tailored to the baseline sample that completed the base questionnaire. When analysing the reading tests, it is therefore advisable to appropriately account for sample non-response and exclusions within the initial sample (according to Table 1 and Table 2). These are mainly due to CQ non-response and, in panel wave 3, to the randomised test administration (see Section 2.5 above). For quite a few of the respondents with missing test data, at least one of two possible repeated test measurements is available (1793 respondents, i.e., 31% of the combined initial samples of panel waves 1 and 3 completed both tests). In addition to randomised test administration, this also makes multiple imputation a promising strategy for handling missing test data in data analyses (for guidance on accounting for test design in imputation, see Section 2.5 and footnote 29).

Table 2: Test response parameters in panel wave 3 (2019)

	Survey language [1)]			Total sample	Non-response and exclusions
	German *N*	French *N*	Italian *n*	*n*	*(%)*
Initial sample with valid base questionnaire (BQ)	4231	1575	348	6154	
of which paper-and-pencil mode	– 192	– 60	– 12	– 264	– 4.3%
Sample with valid BQ data	4039	1515	336	5890	
of which incomplete CQ	– 1158	– 344	– 83	– 1585	– 26.9%
of which valid paper-and-pencil mode of CQ	– 167	– 67	– 16	– 250	– 4.2%
Sample with complete CQ data	2718	1098	239	4055	
of which test not administered due to language[2)]	0	0	– 239	– 239	– 5.9%
of which test randomly not administered by design[3)]	– 672	– 363	0	– 1035	– 25.5%
Initial reading speed test sample	2046	735	0	2781	
of which technical problems with test administration	– 15	– 5	/	– 20	– 0.7%
of which test refused, not valid or missing	– 60	– 16	/	– 70	– 2.5%
Realised sample with valid test data	1971	714	/	2685	

1) Respondents are at liberty to choose one of the three survey languages. With a few exceptions the survey language corresponds to the majority language of the language region. In rare cases, the survey languages of BQ and CQ differ. 2) No test is available for Italian. 3) Testing in one sample split limited to respondents who reported to be in their final year of upper-secondary education in panel wave 3.

For the sake of clarity, it should be mentioned that the missing categories used in Table 2 differ slightly from those in the 2023 data release. Appendix C provides a recoding of the missing categories used there. Note that due to the test administration design there are non-random missing values in the graduation year split-half, which should be taken into account when analysing the entire sample of panel wave 3 (see Section 2.5).

2.5.2 Test-related variables

All datasets comprise the following data:

(tx)rs_score is the sum score of the 51 test items. [32] Only valid tests received a sum score. Invalid tests were marked with a specific missing code. See Appendix B for more information.

(tx)rs_split divides the sample into two randomised halves.
One half undergoes the test in in panel wave 3 unconditionally ("pre-defined wave split-half").
The other half undergoes the test only on condition that respondents are in their final year of upper-secondary education ("graduation year split-half").

(tx)rs_graduationyear a flag indicating whether a respondent is in his or her final year of upper-secondary education in a given panel wave.

(tx)rs_assigned indicates whether a respondent has been assigned to have the test administered in a given panel wave.[33]

For further para-data pertaining to test score analysis and complementary information on the variables listed above, we refer to Appendix B. An overview of the different test datasets and the specific test variables that they comprise is provided in Appendix F.

3 Empirical analyses

The following descriptive analyses only include cases with valid test data from the panel waves published in the TREE2 2023 data release (i.e., panel waves 0 to 3; see TREE, 2023). In Section 3.1 we first explain which cases we have formally classified as non-valid. Section 3.2 contains some descriptive analyses of the test data. Section 3.3 comprises some remarks on the ceiling effects that are to be expected with increasing age of the tested individuals. In the final Section 3.4 we discuss test reliability and validity.

3.1 Formal validity of test data

The adaptation of the test to web-based self-administration (CASI) via smartphone or computer (see Section 2.4.2) has potential implications for the test data's formal validity.

The original paper-and pencil variant of the test is tailored to classroom administration with a test administrator who, among other things, is called to ensure that the time limits are respected. This is prone to create a clearly perceptible test situation — which is not the case with

[32] With the exception of the dataset pertaining to panel wave 2, in which no test was administered.

[33] Note that in wave 3, the test was not assigned to respondents who were in the graduation year split-half, but not in their final year of upper-secondary education (*(tx)rs_assigned* = 0 "no test assigned").

the self-administered screen-based mode, be it at home or on the move. In addition, it cannot be ruled out that individuals other than the respondent complete the test or assist the respondent in doing so. In comparison to the classroom setting, it is also more likely that respondents refuse to complete the test. In the original NEPS test setting, respondents are always supervised. Even in the more recent computer-based version of the test, for instance, trained interviewers accompanied test respondents on the telephone. Incorrect answers are infrequent, as all the test items can be completed without any particular prior knowledge.

Against this background, we have to assume that the adapted test setting affects the frequency of incorrect answers. Skipping items or straightlining due to lacking interest is not unusual in self-administered settings such as CAWI (Zhang & Conrad, 2014). Compared to the corresponding NEPS data, an initial screening of the TREE data shows a higher number of test refusals and of incorrectly solved items. Improper completion of the test thus occurred more frequently than in the NEPS cohort of nearly the same age that completed the paper version (<1% in NEPS compared to 5.7% in TREE2 panel wave 2 resp. 4.4% in panel wave 3).[34]

In view of these results, we have validated the individual tests, among other things, by drawing on para-data with respect to respondents exiting the test prematurely (= test termination), in order to properly and correctly distinguish completed tests from non-valid tests. Non-valid cases are assigned no sum score in the variable "*rs_score*" but are grouped into one of the three categories below. Test results are considered as non-valid if…

- … we have reason to assume that part or all of the test was not properly displayed due to technical problems. The same applies to test results for which no para-data with respect to individual test exit are available [*technical problems*];
- … respondents refused to complete the test or exited it prematurely on their own account before the 2 minutes had elapsed [*refusal/cancellation*];
- … respondents skipped or failed to correctly solve five or more test items (>10% of test non-valid)[35] [*5+ incorrect or skipped/missed answers*].

According to these rules of exclusion, 217 respondents in panel wave 1 and 96 in panel wave 2 were not assigned a valid test score.

[34] It should however be noted that NEPS adopted less restrictive validation criteria than TREE. NEPS has refrained from assessing content validation and, for instance, counted a score of 0 points as valid. Contrary to TREE, NEPS considered, for instance, incomplete tests or test with exclusively erroneous responses as valid.

With respect to the TREE figures, the percentages refer to all non-valid responses of all potential reading test observations. Technical problems are neglected, as external influences made valid testing impossible.

[35] Cross-validations with reading interest, mathematics score, verbal self-concept and grade in teaching language reveal that the correlations between the test score and these validation criteria are maximised when excluding cases who skipped more than 5 tasks or solved them incorrectly.

3.2 Descriptives

Table 3 provides descriptive information on the distribution of test scores in the TREE2 data (TREE, 2023) as well as for the 9th-graders of two NEPS cohorts on which we drew for comparison with panel wave 1 data (10th grade). The table presents unweighted figures as the NEPS statistics were also unweighted. An overview of the weighted descriptives can be found in Appendix D. The results are broken down by test language and by the device respondents used to complete the test.

Table 3: Overview of descriptives (unweighted)

Study & Wave/Cohort	*Grade*	*N (total)*	*N (valid)*	*% missing*	*M*	*SD*	*Reliability*
TREE2 wave 1 (all)	10	4030	3813	5.4	34.2	6.9	0.98
TREE2 wave 1 (German)	10	2949	2792	5.3	34.4	6.9	0.98
TREE2 wave 1 (French)	10	1081	1021	5.6	33.4	6.7	0.99
TREE2 wave 1 (smartphone/tablet)	10	2466	2323	5.8	34.1	6.8	0.98
TREE2 wave 1 (computer/laptop)	10	1564	1490	4.7	34.2	7.0	0.98
TREE2 wave 3 (all)	12	1920*	1854*	3.4	36.2	7.6	0.99
TREE2 wave 3 (German)	12	1382*	1331*	3.7	36.4	7.6	0.99
TREE2 wave 3 (French)	12	538*	523*	2.8	35.7	7.4	0.98
TREE2 wave 3 (smartphone/tablet)	12	1429*	1380*	3.4	36.0	7.5	0.99
TREE2 wave 3 (computer/laptop)	12	491*	474*	3.5	36.8	7.6	0.99
NEPS Starting Cohort 3	9	4898	4888	< 1.0	34.4	8.2	0.98
NEPS Starting Cohort 4	9	14539	14524	< 1.0	34.1	8.9	0.98

*N (total = number of respondents to which the test was administrated; N (valid) = number of respondents with valid test score; % missing = share of respondents without valid test score; SD = standard deviation of test score; Reliability = split-half (even vs. uneven items; missing = false). * N of panel wave 3 refers to pre-defined wave split-half (*rs_split==*1).*

The TREE2 mean values in panel wave 1 (i.e., grade 10) are comparable to those of the NEPS tests (conducted in grade 9). This may seem surprising, as one might expect higher values among the TREE cohort having completed one more grade than the NEPS cohort. We assume that this is due to the fact that written German is not the mother tongue of the Swiss-German sample. From wave 1 to wave 3, we observe a distinct increase of the average score.

T-tests for language differences show small effects and aren't even significant in panel wave 3 (wave 1: $t(1881)= 4.0$, $p < 0.001$, $d= 0.14$; wave 3: $t(972)=1.81$, n. s., $d= 0.09$). Test language effects have already been extensively checked on an item-by-item basis and the validity of translation has been confirmed (see Section 2.4.1 and Appendix A). It therefore remains to be examined here to what extent device differences were found. A T-test of panel wave 1 test data did not show any significant differences with respect to the sum score across the two device-specific translations ($t(3132)=0.32$, n.s.). At the second measurement point (panel wave 3), however, a significant effect

was found for the pre-defined split-half: Those who completed the test with a computer or laptop (M= 36.8; SD=7.6) showed better reading speed performance on average than those using a smartphone or tablet (M= 36.0; SD=7.5). However, the sample of the former is much smaller than that of the latter, and the effect shown can be classified as marginal. ($t(819)= -1.98$, $p < 0.05$; $d= -0.1$).

Overall, the test displays good split-half reliability (Cronbach's alpha = .98) at both panel waves 1 and 3 (for more detailed information on the reliability and validity of the reading speed test scores, see Section 3.4.).

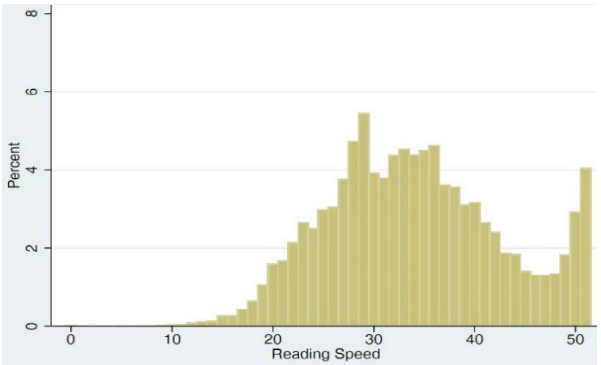
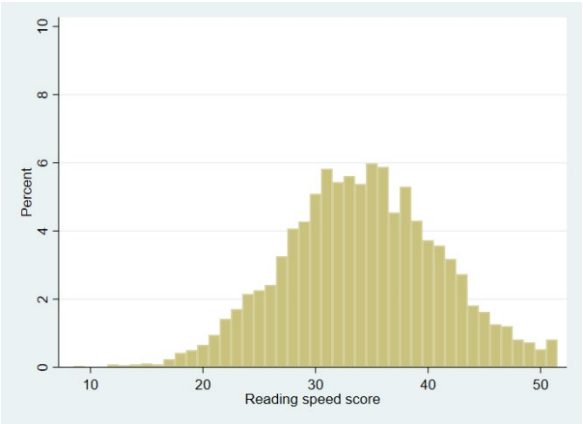
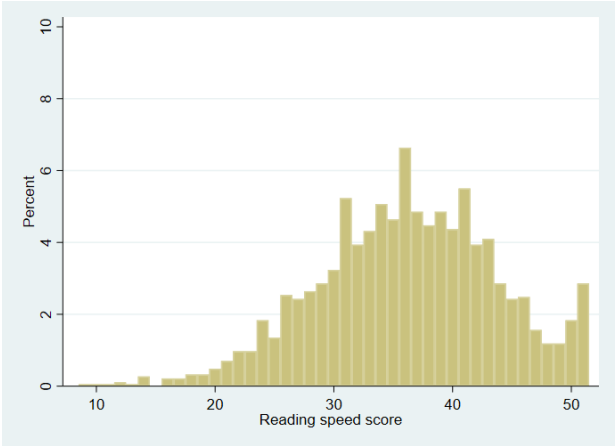
3.3 Ceiling effects

Figure 6 displays the test score distribution for TREE2 panel waves 1 and 3 compared to the 9th-graders of NEPS. In panel wave 1/grade 10 the share of TREE2 respondents who correctly assess all 51 test items within the 2-minute time limit is below 1%. The corresponding share within the NEPS cohort is significantly higher at 4% — despite the fact that they took the test one school year earlier (i.e., in grade 9). However, we must take into account that the TREE2 sample is composed predominantly of German-speaking respondents whose mother tongue is not written German, but Swiss-German dialect. As expected, the ceiling effect increases as the cohort grows older: in panel wave 3 (i.e., grade 12/average age 19), we observe a — still quite modest — ceiling effect of 3% among the TREE2 cohort.

In view of these results, we expect the ceiling effects to further increase as the cohort grows older. Consequently, we expect the test to differentiate less and less well in the upper range of the scale, i.e., among the particularly fast-reading respondents.

As already mentioned, the longitudinal comparability of the test was an important criterion for the instrument to be adopted by TREE. As we are particularly interested in good differentiation in the lower performance range, we consider the ceiling effects observed so far as acceptable — even if they are bound to increase somewhat in the future.

Figure 6: Distribution of test scores: NEPS and TREE2 (panel wave 1 and 3) compared

NEPS («GRADE 9»,2010)	TREE2, WAVE 1 («GRADE 10», 2017)	TREE2, WAVE 3 («GRADE 12», 2019)
		
Mean: 34.1 Median: 33.0 Standard deviation: 8.7 Minimum: 0 Maximum: 51 Ceiling effect for the top 4.0%	Mean: 34.2 [*34.0*] Median: 34.0 [34.0] Standard deviation: 6.9 [*6.9*] Minimum: 9* Maximum: 51 Ceiling effect for the top 0.8% [*0.8%*]	Mean: 36.2 [*35.7*] Median: 36.0 [36.0] Standard deviation: 7.6 [*7.5*] Minimum: 9* Maximum: 51 Ceiling effect for the top 2.9% [*2.9%*]

* *Because of the validation process, a minimum of 0 could not occur.*

[ ] *Numbers in brackets are weighted. Descriptives for wave 3 are calculated by drawing on the pre-defined wave split-half.*

3.4 Test reliability and validity

3.4.1 Split-half & retest reliability

Table 3 shows good split-half reliability (Cronbach's alpha ≥ .98; odd and even items counter-balanced, missing values counted as incorrect) for both languages and both panel waves 1 and 3.

A total of 1793 respondents completed the test in panel waves 1 and 3 and thus provide data on two points of measurement in the TREE2 2023 data release (TREE, 2023). From panel wave 1 to wave 3, we observe a distinct increase of the average score. Retest reliability was determined by means of the correlation between test scores in panel waves 1 and 3 (weighted correlations of all cases with measures in both panel waves: r_{Total}= .80, p<.001; weighted correlations of the pre-defined wave split-half: $r_{split=1}$ = .78, p<.001).[36] The correlation coefficient of .8 indicates that the test measures reading speed reliably. The fact that the correlation is not even higher is not too surprising when keeping in mind that education and training programmes at upper-secondary level in Switzerland vary greatly in terms of total number of lessons taught in a given curriculum - and thus of the extent of training opportunities with respect to reading (speed).

3.4.2 Criterion validity

To assess the criterion validity of the test adaptation, we focus on a selection of known or at least plausible correlates of reading speed. When looking at reading speed in panel wave 1, we observe weak to medium correlations with regard to the following validation criteria that were collected one year before the test (i.e., in the baseline survey):[37] Reading speed (*t1rs_score*) shows a moderate positive correlation with interest in reading (*tointrea*; r_{German} = .32, p < .00 / r_{French} = 0.36, p=.00). The correlation with the verbal self-concept is somewhat weaker (*toscverb_fs*: r_{German} = 0.19, p=.00/ r_{French} = 0.26, p=.00). Grades in teaching language at baseline (*tomarklang*) correlate weakly, yet positively (r_{German} = 0.03, p=.07/ r_{French} = 0.03, p=.00). We also observe a moderate positive correlation with the score of the mathematics test administered at baseline (*towlem*) (r_{German} = 0.34, p=.00/ r_{French} = 0.33, p=.00).

Furthermore, we find a rather weak correlation with the frequency of (pastime) reading surveyed in panel wave 1 (*t1spares*: r_{German_to} = 0.21, p=.00/ r_{French_to} = 0.27, p=.00). Weak to moderate correlations are found for the linguistic self-concept (*t1langself_fs*: r_{German_to} = 0.26, p=.00/ r_{French_to} = 0.31, p=.00).

Overall, the direction and strength of the correlations imply a good criterion validity, as do corresponding validations with test scores from panel wave 3 (see Appendix E). With respect to the paper-and-pencil variant of the test as administered by the NEPS, comparable moderate correlations with the same or similar validation criteria were found: for example, self-concept in the area of reading (r= .31, p < .00) or interest in German language as a school subject (r= .14,

[36] Correlation weighted by the weights of panel wave 3 (t3wt). $r_{split=1}$ refers to the cases in the pre-defined wave split-half.

[37] Correlations weighted by the weights of panel wave 1 (t1wt). For an overview of all weighted and unweighted correlations with respect to both examined panel waves see Appendix E.

p < .00) (for further information see: Lockl et al., 2020). The comparable NEPS findings suggest that the online implementation and the test's translation to French do not affect test validity. The TREE2-related correlations are in part slightly lower than those observed in NEPS, but this may well be due to the lag between the measurement of the validation criteria in the TREE2 baseline survey and the (first) test administration in panel wave 1. When comparing the TREE results with those of the NEPS, we should further take into account that German is not the mother tongue in German-speaking Switzerland.

4 Conclusions and some words of caution

In this paper, we presented the results of a successful implementation and some measurement properties of a bilingual, repeatedly measured reading speed test. The CAWI adaptation and translation of the paper-and-pencil NEPS reading speed test described in this paper has been developed in close consultation with the NEPS team (mainly Karin Gehrer) and in cooperation with the survey institute M.I.S. Trend.[38] Almost six years have passed since its implementation, and in the fast-moving world of digitisation, the technical opportunities for online administration of such tests have much evolved. In the meantime, NEPS has introduced a test extension to limit ceiling effects and has also switched to computer-based tests in 2020.

It should be noted that TREE did not have the opportunity to introduce a control group that completed the test in its "original" NEPS setting (i.e., paper-and-pencil mode in a proctored classroom setting).[39] Therefore, there is no systematic comparison between the survey settings.

When translating the test to French, we have taken great care to keep item-specific length of text as comparable as possible across languages, thus achieving a high measurement equivalence between the two languages. The approach presented is innovative and greatly enhances the analytic potential of the TREE2 data, the more so as test results are available for the majority of the TREE2 sample. Nevertheless, there are slight item-specific differences in reading time that should be considered depending on the research interest. The availability of extensive technical para-data related to the CAWI adaptation allowed us to substantially improve the handling of (formally) non-valid test results. Compared to the original paper-and-pencil test, detailed para-data on test behaviour greatly extend the scope of validation opportunities. Striving for maximum transparency, we provide a detailed Stata syntax [TREE2_Syntax_Reading_Speed_Test_Validation_v2.do] that reveals how we conducted our test validation. Interested scholars thus have the opportunity to make their own adjustments when calculating sum scores.

It should be noted that the paper-to-web adaptation of the test poses certain challenges that may affect test performance. Several studies show that people read more slowly on screen than

[38] https://www.mistrend.ch

[39] This was due to restrictions in terms of implementation deadlines and access to adequate classroom settings.

on paper. Robitzsch and colleagues (2017) report (for Germany) that PISA test items become more difficult on average if they are administered digitally. Since we did not test both modes, we cannot provide a corresponding evaluation. Our test adaptation is therefore likely to be meaningful for digital media reading speed only (Noyes & Garland, 2008; Dillon 1994). These considerations should be taken into account when comparing the TREE test data with those based on paper-and-pencil test administration.

Even though various methodological and content-related issues remain unresolved, the large number of formally flawless tests imply that TREE's digitisation and translation of the test were successful. Plausible distribution patterns of the test results, satisfactory retest reliability and plausible results with respect to external validity further lead us to assume that our adapted instrument provides a reliable and valid measurement of reading speed. Moreover, the results of cross-language validation presented in this paper indicate that the test's French translation provides us with an approximate equivalent measure of reading speed in a second test language.

In conclusion, our digital adaptation, alongside the introduction of a French version of the test substantially extend the scope of our study. The inclusion of the test data in the TREE2 2023 data release provides researchers with new opportunities to investigate reading speed and its role in educational and professional contexts.

5 References

Artelt, C., Stanat, P., Schneider, W., & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In Deutsches PISA-Konsortium (Ed.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 69-137). Leske + Budrich.

Auer, M., Gruber, G., Wimmer, H., & Mayringer, H. (2005). *Salzburger Lese-Screening für die Klassenstufen 5-8*. Hogrefe.

Autorenteam Kompetenzsäule. (2020). Längsschnittliche Kompetenzmessung im NEPS: Anlage und deskriptive Befunde (NEPS Survey Paper No. 80. Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel. https://doi.org/https://doi.org/10.5157/NEPS:SP80:1.0.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Science*. Lawrence Erlbaum.

Hahnel, C., Goldhammer, F., Kröhne, U., Schiepe-Tiska, A., Lüdtke, O., & Nagy, G. (2017). Der Einfluss kognitiver Basisfertigkeiten auf die Änderung der in PISA gemessenen Lesekompetenz. *Zeitschrift für Erziehungswissenschaft*, 1-24. https://doi.org/10.1007/s11618-017-0748-0

Hartmann, S. (2012). *Die Rolle von Leseverständnis und Lesegeschwindigkeit beim Zustandekommen der Leistungen in schriftlichen Tests zur Erfassung naturwissenschaftlicher Kompetenz* Universität Duisburg-Essen, Fakultät für Bildungswissenschaften, Lehrstuhl für Lehr-Lernpsychologie]. Duisburg-Essen. https://d-nb.info/1037311434/34

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. https://doi.org/10.48350/175367

Perfetti, C. A. (1985). *Reading ability*. Oxford University Press.

Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology*, *67*(4), 461–469. https://doi.org/https://doi.org/10.1037/h0077013

Pissarek, M. (2018). Zum Begriff der Lesekompetenz – Förderung von Lesekompetenz bei jungen Erwachsenen. In *Förderung der Lesekompetenz bei Jugendlichen in Ausbildung* (pp. 5-14). Tectum - ein Verlag in der Nomos Verlagsgesellschaft mbH & Co. KG.

Rosebrock, C., & Nix, D. (2006). Forschungsüberblick: Leseflüssigkeit (Fluency) in der amerikanischen Leseforschung und-didaktik. *Didaktik Deutsch*, *20*(2006), 90-112.

Sacchi, S. (2023). Longitudinal Weights for the 2nd TREE-Cohort (TREE2). Construction and Application. Bern: TREE. https://boris.unibe.ch/176648/

Scharfen, J., Blum, D., & Holling, H. (2018). Response Time Reduction Due to Retesting in Mental Speed Tests: A Meta-Analysis. *Journal of Intelligence*, *6*(1), 1-28. https://doi.org/10.3390/jintelligence6010006

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*, 44-66. https://doi.org/10.1016/j.intell.2018.01.003

Sturm, A. (2011). Leseflüssigkeit als Bedingung fürs Textverstehen. *Alfa-Forum*, *76*, 15-17.

Trauzettel-Klosinski, S., & Dietz, K. (2012). Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST. *Investigative Ophthalmology & Visual Science*, *53*(9), 5452-5461. https://doi.org/10.1167/iovs.11-8284

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

Van Buuren, S. (2018). Flexible imputation of missing data. In *Flexible Imputation of Missing Data, Second Edition*. https://www.routledge.com/Flexible-Imputation-of-Missing-Data-Second-Edition/Buuren/p/book/9781032178639

Zhang, C., & Conrad, F. G. (2014). Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127-135.

Zimmermann, S., Artelt, C., & Weinert, S. (2014). National educational panel studies. The assessment of reading speed in adults and first-year students. Bamberg: Leibniz Institute for Educational Trajectories. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC5/3-0-0/com_rs_SC5_SC6.pdf https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC5/3-0-0/com_rs_SC5_SC6.pdf

Zimmermann, S., Gehrer, K., Artelt, C., & Weinert, S. (2012). The Assessment of Reading Speed in Grade 5 and Grade 9. Bamberg: National Educational Panel Study (NEPS). https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/com_rs_2012_en.pdf https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/com_rs_2012_en.pdf

Appendices

Appendix A: T-test tables for reading speed test time differences

T-test tables of the time differences obtained by comparing the German and French versions of the reading speed test.

Table A1: T-test table of language-related time differences (pooled waves 1 and 3)

reading speed items	German			French			(SD1 = SD2)	T-Test (M1=M2)			Cohen's *d* 3)
	M1	*SD1*	*n*	*M2*	*SD2*	*n*	P_1 1)	*t*	df	P_2 2)	
item 1	2.74	1.29	4756	2.85	1.34	1733	.0165	-3.0	2960	.0028	-0.09
item 2	2.51	1.48	4755	2.23	1.00	1736	.0000	8.6	4550	.0000	0.20
item 3	2.49	1.25	4762	2.57	1.29	1732	.0001	-2.2	3003	.0272	-0.06
item 4	2.35	1.00	4760	2.35	1.13	1734	.0000	-0.1	2784	.9392	0.00
item 5	2.08	1.04	4760	1.78	0.81	1734	.0005	12.0	3918	.0000	0.30
item 6	2.96	1.48	4759	2.90	1.23	1735	.9006	1.3	6492	.1833	0.04
item 7	2.86	1.28	4758	2.96	1.32	1733	.0025	-2.8	2993	.0054	-0.08
item 8	2.70	1.27	4760	2.79	1.15	1734	.0456	-2.7	3380	.0064	-0.07
item 9	3.10	1.47	4755	3.18	1.85	1734	.0004	-1.7	2575	.0947	-0.05
item 10	3.20	1.94	4759	3.11	1.40	1733	.4423	1.6	6490	.1005	0.05
item 11	2.25	0.89	4759	2.11	0.85	1735	.5246	5.8	6492	.0000	0.16
item 12	3.11	1.32	4760	3.03	1.40	1733	.0002	2.2	2917	.0310	0.06
item 13	3.39	1.51	4758	3.59	1.67	1732	.0005	-4.2	2825	.0000	-0.12
item 14	3.41	1.45	4741	3.52	1.87	1729	.0000	-2.2	2520	.0265	-0.07
item 15	4.45	2.36	4745	4.26	2.07	1728	.0000	3.2	3464	.0015	0.08
item 16	3.34	1.63	4740	3.19	1.45	1727	.0069	3.4	3417	.0007	0.09
item 17	2.73	1.44	4737	3.02	1.64	1729	.0000	-6.4	2760	.0000	-0.19
item 18	2.98	1.42	4726	3.34	1.73	1724	.0613	-7.6	2614	.0000	-0.23
item 19	4.13	1.93	4712	4.25	1.67	1723	.8366	-2.3	6433	.0235	-0.06
item 20	3.92	1.76	4696	4.41	1.73	1717	.0001	-9.9	3098	.0000	-0.28
item 21	4.45	1.70	4669	4.28	1.51	1709	.0000	3.7	3397	.0002	0.10
item 22	3.39	1.45	4648	3.24	1.30	1694	.0000	3.9	3335	.0001	0.11
item 23	3.74	1.59	4612	4.09	1.51	1674	.0601	-8.1	3117	.0000	-0.22
item 24	3.85	1.68	4527	3.75	1.50	1654	.0308	2.3	3270	.0235	0.06
item 25	3.82	1.41	4445	4.44	1.69	1622	.0000	-13.1	2485	.0000	-0.41
item 26	3.55	1.41	4359	2.96	1.23	1592	.0003	15.8	3200	.0000	0.44
item 27	3.27	1.32	4266	3.57	1.42	1561	.0004	-7.1	2610	.0000	-0.22
item 28	3.13	1.18	4155	3.35	1.29	1512	.0001	-5.8	2485	.0000	-0.18
item 29	3.69	1.44	3996	4.17	1.46	1450	.1772	-10.8	5444	.0000	-0.33
item 30	3.02	1.16	3856	3.16	1.23	1379	.0093	-3.7	2316	.0002	-0.12
item 31	3.06	1.14	3705	3.61	1.27	1300	.0000	-13.8	2068	.0000	-0.47
item 32	3.78	1.51	3483	4.25	1.81	1207	.0000	-8.2	1820	.0000	-0.30
item 33	3.50	1.32	3249	3.82	1.41	1106	.0004	-6.6	1808	.0000	-0.24
item 34	3.36	0.98	3002	3.19	0.88	1028	.0058	5.3	1963	.0000	0.18

reading speed	German			French			*(SD1 = SD2)*	T-Test (M1=M2)			Cohen's *d* 3)
items	*M1*	*SD1*	*n*	*M2*	*SD2*	*n*	P_1 1)	*t*	df	P_2 2)	
item 35	3.14	1.25	2760	3.05	1.05	927	.0264	2.2	1870	.0253	0.08
item 36	3.36	1.02	2513	3.37	0.94	837	.1108	-0.2	3348	.8759	-0.01
item 37	4.08	1.34	2181	3.49	1.05	738	.0000	12.1	1609	.0000	0.46
item 38	3.11	1.28	1963	2.69	0.98	654	.0000	8.7	1443	.0000	0.35
item 39	3.14	0.99	1716	3.24	0.98	575	.2801	-2.2	2289	.0246	-0.11
item 40	3.28	1.14	1500	3.30	0.98	471	.2574	-0.4	1969	.7131	-0.02
item 41	3.32	0.95	1278	3.11	1.03	402	.4373	3.9	1678	.0001	0.22
item 42	3.62	0.98	1039	3.49	1.03	334	.3460	2.0	1371	.0442	0.13
item 43	3.63	0.92	887	3.34	0.80	271	.0419	5.2	506	.0000	0.34
item 44	3.77	1.02	689	3.77	1.07	195	.1997	-0.1	882	.9203	-0.01
item 45	3.02	1.18	556	2.99	0.94	162	.8773	0.3	716	.7963	0.02
item 46	2.80	1.09	431	2.23	0.82	141	.0247	6.6	312	.0000	0.56
item 47	2.98	0.88	351	3.00	0.89	114	.6700	-0.2	463	.8442	-0.02
item 48	3.38	1.22	267	3.02	0.81	93	.0086	3.2	244	.0015	0.32
item 49	3.05	1.07	210	2.71	0.72	68	.0607	2.9	170	.0040	0.34
item 50	2.95	0.87	170	2.92	0.78	59	.2103	0.2	227	.8442	0.03
item 51	3.16	1.19	128	3.68	1.84	48	.0009	-1.8	63	.0765	-0.37

1) Minimal P-Value from three variants of Levene's robust test for equal variances.
2) P-Value for t-test for equal means (assuming equal variances where P1 ≥ .10)
3) Cohen's d *(for groups of unequal size)*

Table A2: T-test table of language-related time differences (panel wave 1)

reading speed items	German			French			(*SD1* = *SD2*)	T-Test (M1=M2)			Cohen's *d* 3)
	M1	*SD1*	*n*	*M2*	*SD2*	*n*	P_1 1)	*t*	df	P_2 2)	
item 1	2.79	1.07	2787	2.97	1.44	1020	.0023	-3.6	1453	.0004	-0.15
item 2	2.59	1.32	2788	2.32	0.98	1022	.0021	6.7	2424	.0000	0.21
item 3	2.56	1.19	2791	2.67	1.27	1019	.0011	-2.4	1709	.0149	-0.09
item 4	2.41	1.00	2789	2.42	1.14	1021	.0007	-0.3	1625	.7910	-0.01
item 5	2.15	1.00	2789	1.86	0.79	1020	.0074	9.1	2267	.0000	0.30
item 6	3.03	1.58	2789	2.99	1.19	1022	.7168	0.8	3809	.4262	0.03
item 7	2.94	1.27	2789	3.05	1.34	1021	.0025	-2.3	1734	.0225	-0.09
item 8	2.80	1.30	2790	2.93	1.25	1020	.0048	-2.7	1884	.0064	-0.10
item 9	3.19	1.44	2787	3.31	1.87	1020	.0004	-1.8	1479	.0770	-0.07
item 10	3.29	2.15	2789	3.22	1.50	1020	.9666	0.9	3807	.3615	0.03
item 11	2.31	0.87	2789	2.18	0.87	1021	.6413	4.2	3808	.0000	0.15
item 12	3.19	1.37	2791	3.12	1.36	1020	.0072	1.3	1818	.2114	0.05
item 13	3.46	1.37	2789	3.69	1.75	1021	.0002	-3.8	1497	.0002	-0.16
item 14	3.51	1.45	2779	3.64	1.99	1019	.0001	-1.8	1431	.0691	-0.08
item 15	4.72	2.55	2784	4.51	2.23	1019	.0006	2.5	2054	.0136	0.08
item 16	3.44	1.71	2782	3.28	1.55	1017	.1544	2.6	3797	.0101	0.09
item 17	2.86	1.58	2779	3.13	1.60	1018	.0011	-4.6	1797	.0000	-0.17
item 18	3.05	1.33	2770	3.44	1.91	1016	.0132	-6.0	1393	.0000	-0.26
item 19	4.29	2.09	2763	4.41	1.69	1016	.6498	-1.6	3777	.1172	-0.06
item 20	3.97	1.75	2752	4.54	1.77	1010	.0001	-8.8	1780	.0000	-0.32
item 21	4.57	1.69	2733	4.35	1.44	1004	.0000	4.0	2080	.0001	0.14
item 22	3.47	1.42	2721	3.33	1.22	995	.0003	2.8	2032	.0044	0.10
item 23	3.85	1.71	2694	4.30	1.61	981	.0779	-7.3	1846	.0000	-0.26
item 24	3.96	1.60	2633	3.87	1.50	963	.0611	1.5	1822	.1370	0.05
item 25	3.95	1.40	2583	4.59	1.78	946	.0007	-9.9	1395	.0000	-0.42
item 26	3.64	1.48	2522	2.99	1.17	923	.0009	13.5	2057	.0000	0.47
item 27	3.33	1.36	2461	3.68	1.46	902	.0037	-6.3	1510	.0000	-0.25
item 28	3.23	1.20	2390	3.45	1.31	866	.0018	-4.4	1424	.0000	-0.18
item 29	3.75	1.45	2277	4.25	1.44	827	.5162	-8.6	3102	.0000	-0.35
item 30	3.09	1.18	2184	3.24	1.27	785	.0156	-2.8	1294	.0051	-0.12
item 31	3.11	1.11	2078	3.63	1.23	731	.0000	-10.1	1175	.0000	-0.46
item 32	3.85	1.48	1941	4.25	1.78	665	.0000	-5.2	997	.0000	-0.26
item 33	3.55	1.21	1797	3.89	1.42	603	.0022	-5.1	914	.0000	-0.26
item 34	3.47	0.98	1627	3.21	0.77	552	.0005	6.3	1188	.0000	0.28
item 35	3.23	1.27	1468	3.11	1.07	484	.0697	2.0	966	.0513	0.09
item 36	3.40	1.03	1324	3.40	0.93	431	.3083	0.1	1753	.9392	0.00
item 37	4.08	1.24	1125	3.53	0.98	371	.0001	8.8	796	.0000	0.47
item 38	3.13	1.29	1006	2.71	0.97	324	.0002	6.3	724	.0000	0.35
item 39	3.17	0.96	843	3.25	0.97	278	.6821	-1.2	1119	.2215	-0.08
item 40	3.31	1.17	731	3.37	0.95	218	.2310	-0.7	947	.5031	-0.05
item 41	3.38	0.89	607	3.19	1.12	185	.7413	2.3	790	.0195	0.20
item 42	3.64	0.91	481	3.55	1.08	147	.3103	1.0	626	.3111	0.10
item 43	3.67	0.92	397	3.41	0.76	120	.0369	3.0	234	.0028	0.29

reading speed	German			French			*(SD1 = SD2)*	T-Test (M1=M2)			Cohen's *d* [3]
items	*M1*	*SD1*	*n*	*M2*	*SD2*	*n*	P_1 [1]	*t*	df	P_2 [2]	
item 44	3.86	1.04	301	3.80	0.91	82	.2836	0.5	381	.6062	0.06
item 45	3.02	0.95	230	2.92	0.85	65	.6807	0.8	293	.4421	0.11
item 46	2.63	0.99	170	2.33	0.90	57	.8993	2.0	225	.0419	0.31
item 47	2.96	0.79	137	3.14	0.94	47	.2307	-1.3	182	.2110	-0.21
item 48	3.42	1.56	98	3.02	0.87	35	.0478	1.8	107	.0683	0.28
item 49	3.00	0.80	72	2.66	0.60	22	.3417	1.8	92	.0726	0.45
item 50	2.89	0.90	54	3.06	1.10	16	.4565	-0.6	68	.5179	-0.19
item 51	3.05	0.85	37	3.44	1.94	13	.0026	-0.7	14	.4880	-0.33

1) Minimal P-Value from three variants of Levene's robust test for equal variances.
2) P-Value for t-test for equal means (assuming equal variances where $P_1 \geq .10$)
3) Cohen's d (for groups of unequal size)

Table A3: t-test table of language-related time differences (panel wave 3)

reading speed items	German			French			(*SD1* = *SD2*)	T-Test (M1=M2)			Cohen's *d* [3)]
	M1	*SD1*	*n*	*M2*	*SD2*	*n*	P_1 [1)]	*t*	df	P_2 [2)]	
item 1	2.67	1.54	1969	2.69	1.18	713	.9296	-0.3	2680	.7832	-0.01
item 2	2.39	1.67	1967	2.10	1.01	714	.0006	5.5	2089	.0000	0.19
item 3	2.39	1.34	1971	2.42	1.30	713	.0422	-0.6	1293	.5685	-0.02
item 4	2.25	1.00	1971	2.24	1.11	713	.0034	0.2	1158	.8196	0.01
item 5	1.99	1.10	1971	1.67	0.83	714	.0095	8.0	1658	.0000	0.31
item 6	2.85	1.31	1970	2.78	1.28	713	.6108	1.2	2681	.2325	0.05
item 7	2.75	1.28	1969	2.83	1.28	712	.1544	-1.6	2679	.1114	-0.07
item 8	2.54	1.21	1970	2.58	0.95	714	.6151	-0.8	2682	.4432	-0.03
item 9	2.97	1.50	1968	3.01	1.79	714	.1834	-0.5	2680	.6128	-0.02
item 10	3.07	1.59	1970	2.96	1.23	713	.1011	1.7	2681	.0974	0.07
item 11	2.16	0.91	1970	2.00	0.80	714	.2025	4.0	2682	.0001	0.18
item 12	3.01	1.23	1969	2.89	1.44	713	.0139	1.9	1111	.0594	0.09
item 13	3.30	1.68	1969	3.44	1.52	711	.1974	-1.9	2678	.0539	-0.08
item 14	3.26	1.44	1962	3.35	1.68	710	.0255	-1.3	1107	.2094	-0.06
item 15	4.07	1.99	1961	3.90	1.75	709	.0039	2.1	1412	.0335	0.09
item 16	3.20	1.49	1958	3.07	1.28	710	.0062	2.1	1450	.0335	0.09
item 17	2.55	1.19	1958	2.86	1.69	711	.0000	-4.5	977	.0000	-0.23
item 18	2.89	1.52	1956	3.19	1.42	708	.8149	-4.6	2662	.0000	-0.20
item 19	3.90	1.66	1949	4.02	1.63	707	.9475	-1.7	2654	.0904	-0.07
item 20	3.85	1.77	1944	4.22	1.66	707	.2314	-4.8	2649	.0000	-0.21
item 21	4.27	1.70	1936	4.19	1.60	705	.1330	1.1	2639	.2899	0.05
item 22	3.28	1.49	1927	3.11	1.39	699	.0047	2.7	1320	.0061	0.12
item 23	3.58	1.39	1918	3.80	1.31	693	.8518	-3.7	2609	.0003	-0.16
item 24	3.70	1.77	1894	3.58	1.48	691	.2381	1.6	2583	.1042	0.07
item 25	3.65	1.40	1862	4.23	1.53	676	.0057	-8.7	1107	.0000	-0.41
item 26	3.43	1.29	1837	2.92	1.31	669	.0818	8.6	1168	.0000	0.39
item 27	3.19	1.26	1805	3.41	1.34	659	.0603	-3.6	1105	.0004	-0.17
item 28	3.00	1.15	1765	3.22	1.27	646	.0099	-3.8	1058	.0001	-0.18
item 29	3.62	1.43	1719	4.06	1.48	623	.1928	-6.5	2340	.0000	-0.31
item 30	2.93	1.14	1672	3.07	1.16	594	.2446	-2.5	2264	.0135	-0.12
item 31	3.01	1.16	1627	3.59	1.32	569	.0002	-9.4	892	.0000	-0.49
item 32	3.68	1.55	1542	4.26	1.86	542	.0010	-6.4	820	.0000	-0.35
item 33	3.43	1.43	1452	3.73	1.39	503	.0388	-4.2	899	.0000	-0.21
item 34	3.24	0.96	1375	3.17	0.98	476	.9193	1.4	1849	.1554	0.08
item 35	3.04	1.22	1292	2.98	1.03	443	.2796	1.0	1733	.2994	0.06
item 36	3.32	1.02	1189	3.33	0.95	406	.2334	-0.3	1593	.7346	-0.02
item 37	4.07	1.43	1056	3.45	1.12	367	.0000	8.4	813	.0000	0.45
item 38	3.08	1.26	957	2.67	1.00	330	.0002	6.0	716	.0000	0.34
item 39	3.10	1.02	873	3.23	0.99	297	.2674	-1.9	1168	.0520	-0.13
item 40	3.24	1.10	769	3.24	1.00	253	.6552	0.1	1020	.9272	0.01
item 41	3.27	1.00	671	3.03	0.94	217	.4488	3.1	886	.0022	0.24
item 42	3.60	1.03	558	3.45	0.98	187	.7311	1.7	743	.0810	0.15
item 43	3.61	0.92	490	3.27	0.83	151	.3282	4.0	639	.0001	0.38

reading speed	German			French			*(SD1 = SD2)*	T-Test (M1=M2)			Cohen's *d* [3]
items	*M1*	*SD1*	*n*	*M2*	*SD2*	*n*	P_1 [1]	*t*	df	P_2 [2]	
item 44	3.69	1.00	388	3.76	1.18	113	.0121	-0.5	161	.5853	-0.06
item 45	3.02	1.32	326	3.04	1.00	97	.7386	-0.2	421	.8698	-0.02
item 46	2.91	1.13	261	2.16	0.76	84	.0069	6.9	209	.0000	0.72
item 47	2.99	0.94	214	2.90	0.85	67	.6695	0.7	279	.4814	0.10
item 48	3.36	0.99	169	3.02	0.78	58	.0690	2.7	125	.0085	0.36
item 49	3.07	1.19	138	2.74	0.77	46	.0915	2.2	119	.0305	0.30
item 50	2.98	0.86	116	2.87	0.62	43	.0256	0.9	104	.3952	0.13
item 51	3.21	1.31	91	3.76	1.82	35	.0319	-1.6	48	.1066	-0.38

1) Minimal P-Value from three variants of Levene's robust test for equal variances.
2) P-Value for t-test for equal means (assuming equal variances where $P_1 \geq .10$)
3) Cohen's d (for groups of unequal size)

Appendix B: Overview of variable names and labels, value labels and annotations across datasets

This Appendix provides an overview of the test-related variables and para-data.

Variable name	*Variable label*	*Value labels*
rs_score	Reading speed score	-919 / .s not administered [in graduation year in wave 2] -925 / x. invalid answer [technical problem] -926 / y. invalid answer [within-test break-off, no answer] -927 / .z invalid answer [5+ items skipped/ incorrect] 0 - 51
Annotation rs_score: This is the sum score of the 51 reading speed test items. Formally non-valid test data are those that either had technical problems, exited the test prematurely or did not complete it, or subjects who assessed more than 5 test items incorrectly or skipped them. Formally non-valid test data are not assigned a sum score in this variable but may be analysed in more detail by drawing on the dataset '*TREE2_Data_Reading_Speed_Test_Items_v2*', and scores can be calculated if necessary. Depending on the dataset, the number of specific missing categories varies.		
rs_split	Sample split: Timing of reading speed test administration	-920 / .n not administered [sample split design] 0 graduation year split-half 1 pre-defined wave split-half
Annotation rs_split: The variable divides the TREE2 sample into two randomised halves in panel wave 3. 277 cases were not in the initial sample at this point (refusal to participate) and were not taken into account in the formation of the split (marked with -920 /.n not administered [sample split design]). Cases assigned to the "pre-defined" wave split-half are administered the test unconditionally in panel waves 1 and 3. Cases assigned to the "graduation year" split-half are administered the test only if they are in their final year of upper-secondary education or training (*rs_graduationyear* = 1).		
rs_graduationyear	Graduation year flag (used for reading speed test administration)	0 not in graduation year 1 in graduation year
Annotation rs_graduationyear: This variable is a flag to determine whether a respondent is in his or her final year of upper-secondary education. It is based on an item where respondents could specify the year in which they expected to complete their education. Along with the variable *rs_split*, this variable is used to implement the test sampling design (i.e., to which respondents the test is assigned). For 138 cases in panel wave 2, the flag was constructed on grounds of other data pertaining to educational status and progress.		

rs_assigned	Reading speed test assignment	0 no test assigned 1 test assigned 2 test erroneously assigned
Annotation rs_assigned: The assignment variable indicates when a test should be administered according to the "randomised test administration scheme". Since the split-half design was only implemented from wave 3, the data for waves 1 and 2 are constructed retrospectively: In wave 1 the test was administered unconditionally and thus all cases are assigned *(rs_assigned* = 1 "test assigned"). In wave 2, 69 cases that would have been in the graduation year split-half are also in the last year of upper-secondary education and thus also constructed as assigned (*rs_assigned* = 1). In wave 3, the randomised test administration takes effect and all cases in the predefined wave split-half and the graduates of the graduation year split-half receive an assigned flag (*rs_assigned* = 1). In addition, there are six cases in wave 3 that were erroneously assigned a reading test (*rs_assigned*=2 "test erroneously assigned"). They received the test due to incorrect filtering, although they were not in upper-secondary education.		

Test-related para-data that are only comprised in the dataset '*TREE2_Data_rs_v2*'

Variable name	*Variable label*	*Value labels*
rs_status:	Status of reading speed test	0 Valid test 1 Technical problem 2 Incomplete test (break-off, no answer) 3 Gave 5 or more wrong answers 4 Skipped 5 or more items
Annotation rs_status: This variable specifies the categories of validation. For more information on validation of test scores, see Section 3.1		
rs_correct	Number of correct answers	0 – 51
rs_wrong	Number of wrong answers	0 - 32
rs_lastitm	Index of last completed item	0 - 51
rs_testcomments	Case identification by comments (technical problems or break-off)	0 Comment not test-related 1 Technical problem comment 2 Break off comment
rs_intro	Intro page shown	0 Intro page not shown 1 Intro page shown
rs_testpage	Test page shown	0 No test page shown 1 At least 1 test page shown
rs_timeisup	Time-is-up page shown	0 Time-is-up page not shown 1 Time-is-up page shown

Variable name	*Variable label*	*Value labels*
rs_end	End page shown	0 End page not shown 1 End page shown
rs_exit	Exit-before-time-is-up page shown	0 Exit page not shown 1 Exit page shown
rs_comm	Comments page shown (last survey page)	0 Comments page not shown 1 Comments page shown
rs_lastpage	Position at which test ended	1 Before the test 2 At intro page of test 3 After the test
rs_item_1	Test item 1	0 Wrong answer 1 Correct answer
...	...	...
rs_item_51	Test item 51	0 Wrong answer 1 Correct answer

Appendix C: Assignment of missing value codes and sample statistics in Stata and SPSS dataset

Table A4: Missing information

	Missing information			Total sample	Non-response and exclusions
	SPSS	*STATA*	*Missing label*	*n*	*(%)*
Initial sample with valid base questionnaire (BQ)				6154	
of which paper-and-pencil mode (test not administered)	-922*	.t	Not administered [no BQ / CATI]	– 264	– 4.3%
Sample with valid BQ data				5890	
of which incomplete complementary questionnaire (CQ)	-923*	.u	Not administered [no CQ]	– 1585	– 26.9%
of which valid paper-and-pencil mode of CQ (test not administered)	-905	.i	Not administered [mode]	– 250	– 4.2%
Sample with complete CQ data				4055	
of which test not administered due to language (Italian)	-917	.q	Not administered [CQ not German/French]	– 239	– 5.9%
of which test not administered due to test sampling design	-921	.l	Not administered [sample split design - not in graduation year]	– 1035	– 25.5%
Initial reading speed test sample				2781	
of which technical problems with test administration	-925	.x	Invalid [technical problem]	– 20	– 0.7%
	-926	.y	Invalid [within test break-off, no answer]		
of which test refused, not valid or missing	-927	.z	Invalid [5+ items skipped/incorrect]	– 70	– 2.5%
Realised sample with valid test data				2685	

* *In the wave-specific datasets of the TREE2 2023 data release (TREE, 2023), the missing values are further specified:*

-922 (.t) is further differentiated into -901 (.e) «Not administered [wave]» and -911 (k) «No response [wave]».

-923 (.u) is further differentiated into -906 (.e) «Not administered [survey part]» und -912 (k) «No answer [survey part]».

Appendix D: Weighted descriptives of Table 3 (page 24)

Table A5: Overview of descriptives (weighted)

Study & Wave	Grade	N (total)	N (valid)	% missing	M	SD	Reliability
TREE2 T1 (all)	10	4030	3813	5.4	34.0	6.9	0.98
TREE2 T1 (German)	10	2949	2792	5.3	34.2	6.9	0.98
TREE2 T1 (French)	10	1081	1021	5.6	33.5	7.0	0.98
TREE2 T1 (phone/tablet)	10	2466	2323	5.8	33.8	6.9	0.98
TREE2 T1 (computer/ laptop)	10	1564	1490	4.7	34.3	6.9	0.98
TREE2 T3 (all)	12	1920*	1854*	3.4	35.7	7.5	0.99
TREE2 T3 (German)	12	1382*	1331*	3.7	36.1	7.6	0.99
TREE2 T3 (French)	12	538*	523*	2.8	36.0	7.2	0.98
TREE2 T3 (phone/tablet)	12	1429*	1380*	3.4	35.3	7.5	0.99
TREE2 T3 (computer/ laptop)	12	491*	474*	3.5	36.4	7.7	0.99

Notes:
N (total) = Number of respondents to whom the test was administered (unweighted)
N (valid) = Number of respondents with valid test score (unweighted)
% missing = Share of respondents without valid test score (unweighted)
M = Weighted mean of test score,
SD = Weighted standard deviation of test score
Reliability = Weighted split-half-reliability (even vs. uneven items; missing values = erroneous assessment of item).

Appendix E: Correlations with validation criteria

Table A6: Correlations with validation criteria, panel wave 1

	Unweighted							Weighted					
Variables	German		French		Total			German		French		Total	
	corr	(sig.)	corr	(sig.)	corr	(sig.)	n	corr	(sig.)	corr	(sig.)	corr	(sig.)
t0marklang1	0.03		0.02		0.03		3765	0.03	+	0.03	**	0.03	***
t0scverb_fs	0.21	***	0.25	***	0.22	***	3801	0.19	***	0.26	***	0.21	***
t0intrea_fs	0.30	***	0.32	***	0.30	***	3801	0.32	***	0.36	***	0.33	***
t0wlem	0.31	***	0.32	***	0.31	***	3785	0.34	***	0.33	***	0.33	***
t1spare5*	0.21	***	0.25	***	0.22	***	3812	0.21	***	0.27	***	0.22	***
t1langself_fs	0.28	***	0.28	***	0.28	***	3812	0.26	***	0.31	***	0.28	***

* *The item t1spare5 was reverse-coded for correlation analysis, such that high scores now correspond to a high value of the variable.*

Table A7: Correlations with validation criteria, panel wave 3

	Unweighted							Weighted					
Variables	German		French		Total			German		French		Total	
	corr	(sig.)	corr	(sig.)	corr	(sig.)	n	corr	(sig.)	corr	(sig.)	corr	(sig.)
t0marklang1	0.01	n.s.	0.09	*	0.05	+	2659	0.03	n.s.	0.10	*	0.06	*
t0scverb_fs	0.22	***	0.17	***	0.21	***	2680	0.20	***	0.17	***	0.19	***
t0intrea_fs	0.25	***	0.34	***	0.28	***	2679	0.26	***	0.34	***	0.28	***
t0wlem	0.37	***	0.31	***	0.35	***	2659	0.40	***	0.20	***	0.34	***
t1spare5*	0.19	***	0.25	***	0.21	***	2301	0.19	***	0.24	***	0.20	***
t1langself_fs	0.25	***	0.23	***	0.25	***	2309	0.26	***	0.22	***	0.25	***

* *The item t1spare5 was reverse-coded for correlation analysis, such that high scores now correspond to a high value of the variable.*

Appendix F: Data structure

This chapter explains the organisation of the test data. The data are available in several datasets that serve different analysis strategies. This is due to the complex study and test design that results in a partially selective test sample. In the three Figures below we attempt to visualise which cases and data are contained in the respective data sets.

To understand the data structure, it is important to know that at the time of panel wave 3, the data was divided into two split-halves. One half completed the test at pre-defined panel waves (i.e., panel waves 1 and 3; blue rectangles in the Figures). The other half completed the test on condition that respondents were in their final year of upper-secondary level education (green rectangles). For detailed information on the longitudinal test design, we refer to chapter 2.5.[40]

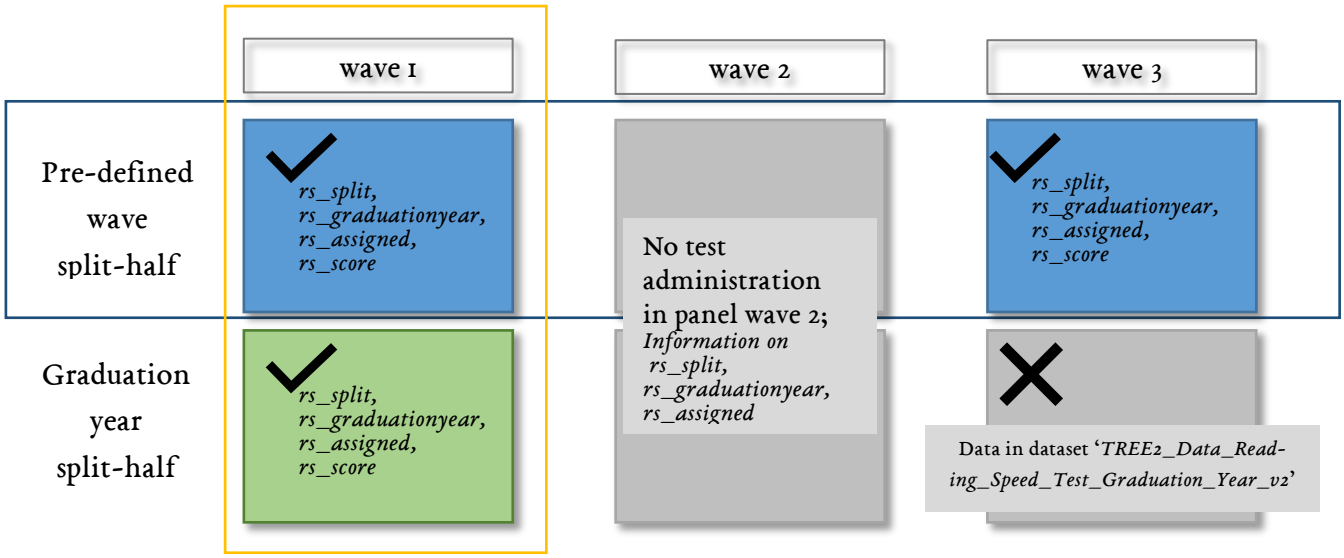
1) Panel wave-specific datasets ('TREE2_Data_Wave_1_v2' and 'TREE2_Data_Wave_3_v2'[40, 41]

The datasets pertaining to panel waves 1 and 3 comprise para-data and test scores that may be drawn on for analyses of intra-individual change or the stability of basic reading skills. For these purposes, scholars should draw on the data from the "pre-defined wave split-half" (outlined in blue in Figure 7). Moreover, the data from panel wave 1 is available for the entire sample and may be used for any analyses, e.g., comparisons between the two TREE cohorts (outlined in orange in Figure 7).

[40] A note pertaining to panel wave 2: Since the randomised test administration design was only implemented from panel wave 3 onwards, no test was administered in wave 2. In total, 138 respondents in panel wave 2 were in 2-year VET programmes and hence were in their last year of training. We were able to identify these cases retrospectively on grounds of training information they reported. In the datasets of the 2023 data release, the variable of the graduation year (*rs_graduationyear*) was therefore set to 1 for these cases and the test score variable (*rs_score*) contains the information that these cases are "not administered [in graduation year in wave 2]". In addition, all of the aforementioned datasets contain para-data for these missed wave 2 cases (rs_assigned, rs_graduationyear, rs_split). However, only the dataset of panel wave 2 includes complete para-data for all relevant cases. The additional data sets related to the reading speed test always contain only a partial sample.

[41] File names of datasets according to the 2023 data release (TREE, 2023).

Figure 7: Test data in panel wave-specific datasets from panel waves 1 and 3

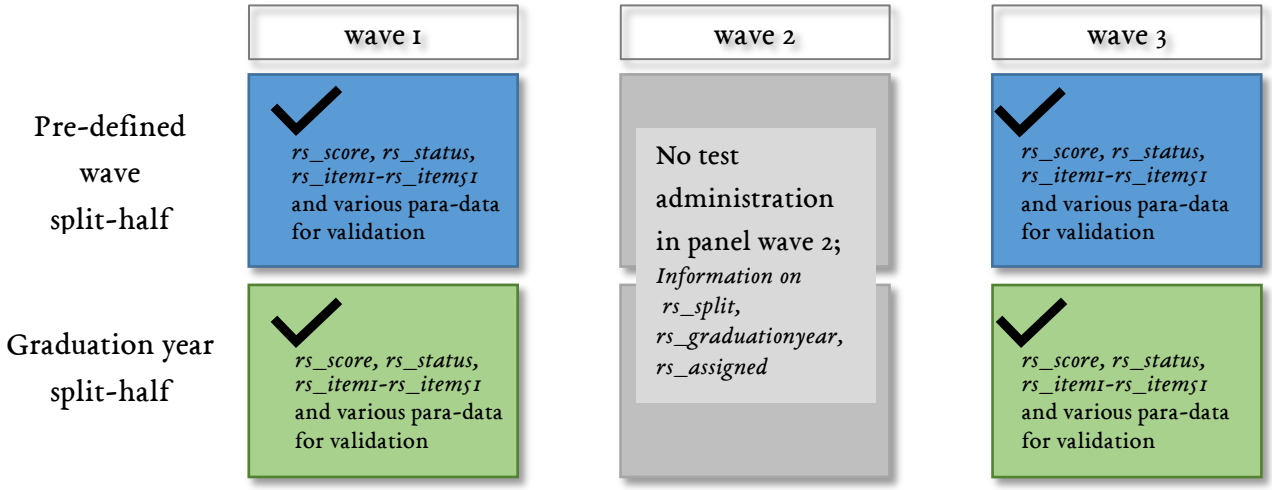


2) *Reading speed test dataset including all test cases ('TREE2_Data_Reading_Speed_Test_Items_v2')*[40, 41]

This dataset includes the cases who completed the complementary questionnaire at least up to the introduction page of the test (see Figure 8; n T1 = 4030 / n T3 = 2781 / n T1 and T3 = 1793). It contains all the data required to calculate and validate the test scores. Therefore, the sample in this data set is smaller than in the panel-wave-related datasets described in 1) above.

The dataset is kept in long format and contains the data of the two tests administered in panel waves 1 and 3.[40] The test data folder in the TREE2 2023 data release (TREE, 2023) also comprises a Stata script that shows how test scores and validations were calculated.

Figure 8: Test data in dataset including all test cases [reading speed test items v2]



3) Reading speed test dataset, "graduation year" sample only
('TREE2_Data_Reading_Speed_Test_Graduation_Year_v2)[40, 41]

This dataset includes the test scores and selected para-data for the split-half sample to which the test was administered specifically in the last year of upper-secondary education and training (graduation year split-half: Figure 9). For possible imputation purposes, the dataset includes not only the split data from wave 3 onwards but also the data of the test administered (unconditionally) to this split-half in panel wave 1 (scores in wave 1 [*rs_score*] as well as selected para-data on graduation year in wave 2[29]).

Figure 9: Test data in dataset [*reading speed test graduation year v2*]