uni.lu
UNIVERSITÉ DU
LUXEMBOURG

LE GOUVERNEMENT
DU GRAND-DUCHÉ DE LUXEMBOURG
Ministère de l'Éducation nationale
et de la Formation professionnelle

épreuves standardisées

ÉpStan Technical Report

www.epstan.lu

Editorship: Antoine Fischbach, Sonja Ugen, & Romain Martin

Workgroup (alphabetical): Rachid Boualam, Christophe Dierendonck, Antoine Fischbach, Eric François, Sylvie Gamo, Thierry Heck, Danielle Hoffmann, Caroline Hornung, Ulrich Keller, Willibrord Koch, Dalia Lorphelin, Romain Martin, Claire Muller, Monique Reichert, Markus Scherer, Philipp Sonnleitner, Sonja Ugen, Denise Villányi, Gina Wrobel

ISBN: 978-2-919940-10-3

Copyright: University of Luxembourg, ECCS research unit/LUCET

Luxembourg, 2014

URL: http://www.epstan.lu

Contents

1 Item Development and Test Compilation

Philipp Sonnleitner, Monique Reichert, Sonja Ugen

1.1 Item Development

1.1.1 Conceptual Background

The major aim of the *Épreuves Standardisées* (ÉpStan) is to evaluate the national school system; that is, to provide information about whether students from primary and secondary schools have achieved key competencies at the level required for their successful participation in subsequent learning contexts—be it in their daily private or school-related lives.

Those key competencies are described in the so-called *Bildungsstandards* (educational standards), which are specified by the *Luxembourg Ministry of Education and Vocational Training* (MENFP) for the different school years (cycles) and subjects. The items that were developed for the ÉpStan are therefore linked to those educational standards and provide information about whether students from Luxembourgish schools have attained particular levels. For the item development process, this means that the items are thoroughly verified with respect to their link to the educational standards (in terms of their content and their difficulty).

The ÉpStan focuses on competencies that ensure an economic and highly standardized assessment. For this reason, all items that are developed are either in a closed format (i.e., multiple-choice, true-false, ordering or matching items) or they require short answers only. It should be noted that not all the competencies mentioned in the educational standards for a given subject are assessed by the ÉpStan: productive language skills, for instance, are not part of the evaluation. Including the evaluation of productive skills such as writing would indeed mean a considerable increase in the work load with respect to the elaboration of scoring criteria, the training of expert raters, and the (time-consuming) rating process itself (which requires an important number of expert raters). Currently, the ÉpStan competency tests are administered at the beginning of grade 1 (cycle 2.1), grade 3 (cycle 3.1), and grade 9 to assess the previous learning cycle (which typically corresponds to 2 school years). In grade 1, the ÉpStan assesses mathematics, Luxembourgish listening comprehension, and early literacy skills (in Luxembourgish). The tests in grade 3 encompass mathematics, German listening comprehension, and German reading comprehension. In grade 9, the tests include mathematics as well as German and French reading comprehension. Further, students and/or parents (depending on the grade) are asked to complete a questionnaire on their socio-economic and linguistic backgrounds—note that this is also the case in grade 7.

1.1.2 General Procedure and Organization

The item pool that is used for the compilation of the ÉpStan tests is developed in annual item-development cycles. The item development that was used to prepare the main ÉpStan test of the school year 2013-2014 will serve as a prototypical example. The item development begins a year prior to the main test (in autumn 2012) and lasts for a complete school year (until July 2013). To ensure strong test quality, items for each grade and subject are developed in teams consisting of researchers (members of the ÉpStan team), teachers actively teaching in the respective grades, and members of the MENFP. In this way, each interdisciplinary team regroups experts in the domains of psychometrics, test development, subject contents, the educational curriculum, and the reference documents.

The teams meet at regular intervals throughout the item-development process. At the beginning of the process, new team members receive training in item development (theory, processes, formats), information about the test design, and information about the ÉpStan in general. During the meetings, newly developed items are thoroughly discussed with respect to the target competency, difficulty level, adequacy with regard to the target population, and psychometric quality (e.g., standardized response format, unambiguous correct answer, or clearly formulated task). Items are modified until agreement about the items' contents and their characteristics is achieved between all team members.

The finalized items are then pretested (in May[1]) to each grade. To evaluate the pretest items, the same psychometric criteria are applied as were applied for the main test[2] (see Section 2.2). Members of the ÉpStan team are present during the pretests in the schools to observe the reactions of the students as they take the tests and to monitor aspects such as the time needed to complete the items. After the pretest, the item development teams re-discuss the items on the basis of the empirical findings. Items with low psychometric qualities are analyzed in depth in terms of their contents and wording, the proposed correct responses, the distractors, and the items' adequacy for the target population. Other adjustments concerning layout or timing issues (esp. for the tests that are administered via a CD) are discussed. If necessary, stimuli contents and/or item contents or their defining characteristics are revised or modified. Finalized items and their characteristics are introduced into the ÉpStan item database. The next step is the main test compilation described in Section 1.2.

Parallel to the item-development meetings, the ÉpStan team members who are responsible for item development meet on a regular basis in the so-called *Item*

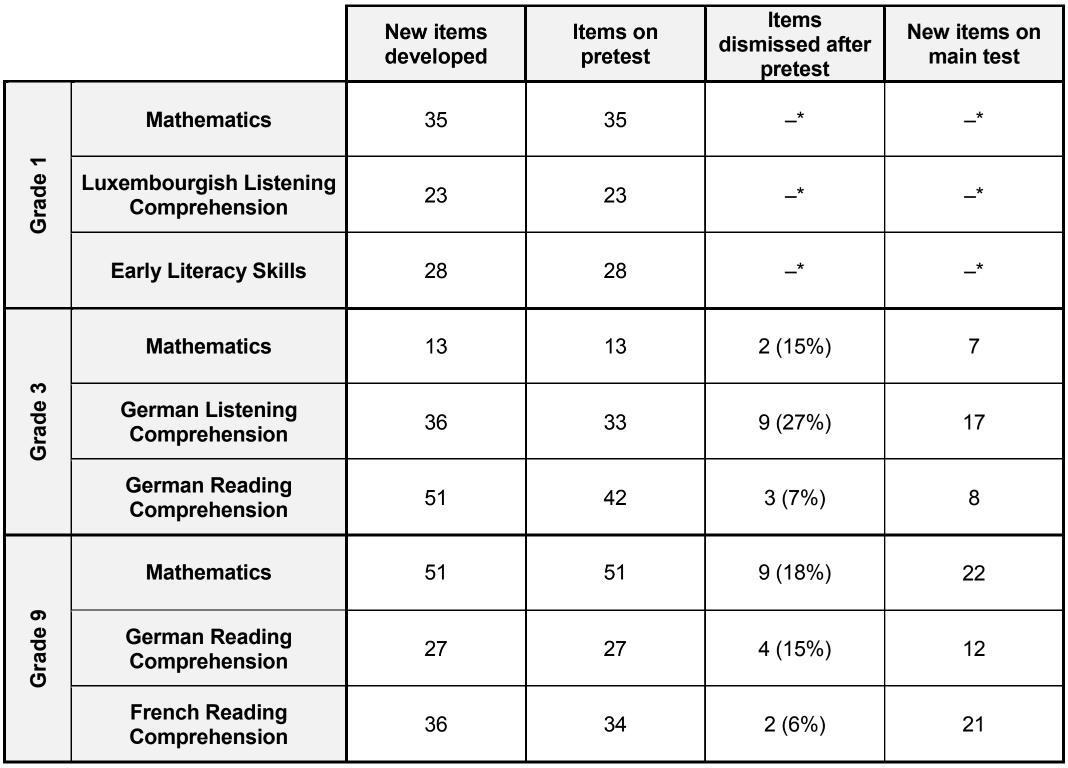
[1] Due to the steep developmental curve in grade 1, items for grade 1 are pretested at a different time in the item-development cycle.

[2] On the pretest, the statistical procedures concerning the longitudinal anchoring are not taken into account.

Development Council (IDC) to discuss the item-development processes, to peer-review the item contents, to examine the results, and to discuss scientific literature on item development. This process ensures that the different ÉpStan tests are developed according to the same quality standards.

1.1.3 Specifics of Item Development

Table 1 shows an overview of the number of items that were developed and pretested during the item-development cycle 2012-2013 per grade and per subject. The numbers of items dropped after the pretest were within acceptable ranges especially since, for most of these items, the cause of the problem could be identified and removed.



		New items developed	Items on pretest	Items dismissed after pretest	New items on main test
Grade 1	Mathematics	35	35	–*	–*
	Luxembourgish Listening Comprehension	23	23	–*	–*
	Early Literacy Skills	28	28	–*	–*
Grade 3	Mathematics	13	13	2 (15%)	7
	German Listening Comprehension	36	33	9 (27%)	17
	German Reading Comprehension	51	42	3 (7%)	8
Grade 9	Mathematics	51	51	9 (18%)	22
	German Reading Comprehension	27	27	4 (15%)	12
	French Reading Comprehension	36	34	2 (6%)	21

Table 1. Overview of the numbers of developed and pretested items during the item-development cycle 2012-2013 with respect to subjects and grades. *not yet applicable

1.1.3.1 Grade 1

Item development for grades 3 and 9 has followed the procedures described above since 2009. From 2014-2015 on, the ÉpStan will be extended to grade 1 to evaluate what the children have accomplished in preschool (at the beginning of grade 1). Therefore, grade-1 items were developed for the first time in 2012-2013 and pretested at the beginning of 2013-2014. Administering group tests at this young age presents an additional challenge. To select age-adapted item types that can be administered as a standardized group test, several small-scale and informal pilot studies were conducted. Compared to the other grades, the timeframe for the pretest was also adapted due to the highly dynamic cognitive development of this age group. Indeed, in grades 3 and 9, the pretest takes place at the end of the school year. The pretest in grade 1, however, will be conducted at the beginning of grade 1, right after the main test (from 2014-2015 onwards).

The grade-1 item-development groups developed tests for the domains of mathematics, early literacy skills (in Luxembourgish), and Luxembourgish listening comprehension based on the reference document published by the MENFP (2010). Special attention was placed on developing mostly language-free instructions and on presenting the items in an appealing way. Further, CDs were developed to present the language-test instructions and items to guarantee a highly standardized test administration.

1.1.3.2 Grade 3

The grade-3 item-development groups developed tests for the domains of mathematics, German listening comprehension, and German reading comprehension based on the reference documents provided by the MENFP (2008a, 2010).

For mathematics, item development covers two content domains: (a) numbers and operations, and (b) space and shape. Item development further covers two contexts (applied vs. not applied). To keep the influence of reading comprehension at an absolute minimum, special attention was paid to keeping the items mostly language-free.

For German reading comprehension as well as German listening comprehension, items mainly address two sub-competencies: (a) locating and understanding (text-based) information, and (b) interpreting written/oral information and applying reading/ listening strategies. For the listening comprehension test, a CD including the stimuli, items, and test instructions was also developed.

1.1.3.3 Grade 9

The grade-9 item-development groups developed tests for the domains of mathematics, German reading comprehension, and French reading comprehension.

The items developed for mathematics tackle four different domains (numbers and operations, geometry, algebra, data) that are described in the reference document provided by the MENFP (2008b, 2008b). Each of the developed items is qualified with regard to the domain and the sub-competency it measures as well as with regard to its difficulty by taking into account characteristics such as the minimum number of calculations to be performed or the number of numerical units the student must integrate to successfully solve the item.

As for grade 3, the developed items contain only a minimum amount of text. Nevertheless, the subject of mathematics—although taught in German in primary schools—is taught in French from the beginning of secondary school on, and students might feel more or less able to solve the items from this domain in either French or German. Therefore, the items are prepared in both German and French. During the test, students can choose the test language for each item and also have access to both versions. However, implementing this choice involves an additional work step for this group to thoroughly verify the translation equivalences of the mathematical items.

In both French and German, reading comprehension items mainly address 2 sub-competencies (a) locating text-based information and globally classifying texts and (b) analyzing and interpreting texts. The latter were deduced from the reference document provided by the MENFP (Kühn, 2008). The stimuli include continuous and discontinuous texts (at a ratio of 2 to 1).

1.1.4 Pretest Sample

In grade 1, 505 students from 32 classes participated in the pretest of the three tests that were newly developed.[3]

In grade 3, 276 students from 20 classes worked on the mathematics items, 240 students from 19 classes worked on the German reading comprehension items, and 255 students from 17 classes worked on the German listening comprehension items. Overall, a representative sample of students participated in the pretest, leading to a valid impression of the functionality of the newly developed items for grade 3.

In grade 9, 263 total students participated in the pretesting of the mathematics items from the three different school forms (ES: 116; EST: 59; EST-PREP: 88).[4] For the German reading comprehension pretest, 260 students participated overall (ES: 110; EST: 66; EST-PREP: 84). For the pretest on French reading comprehension, 259 students participated overall (ES: 108; EST: 60; EST-PREP: 91).

Note that in grade 9, the ratio of students from the three school forms who took the pretest was not perfectly comparable to the ratio of students from those school forms within the complete Luxembourg school system. Subsequent discussions of the items' difficulty estimates took into account this weakness with regard to the sample's representativeness.

1.2 Test Compilation

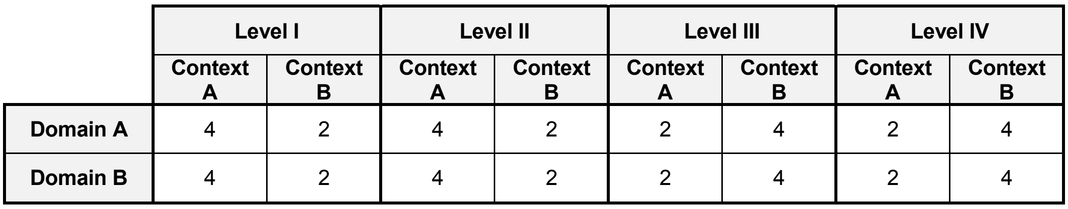
1.2.1 General Procedure

Independent of grade and subject, the test compilation for the ÉpStan main test follows two steps. The first step (1) includes selecting from the existing item pool a certain number of appropriate items that fit a given statistical design for each test (see Table 2

[3] As the pretest had just taken place when this document was submitted, we were not able report any information about the psychometric qualities of the grade-1 pretest items here.

[4] ES: enseignement secondaire; EST: enseignement secondaire technique théorique et polyvalent; EST-PREP: enseignement secondaire technique préparatoire et pratique.

for an example). For each of the ÉpStan tests, a specific test design was developed on the basis of the theoretical reference documents by the MENFP (see Section 1.1.3). This design guarantees that (a) for each subject, all tested competencies are equally represented on the test, (b) each test includes the entire range of theoretical difficulties, and (c) a sufficient number of items that were already administered in previous years is included. Especially the last criterion guarantees that a central aim of the ÉpStan—i.e., the comparison of the performance of different student cohorts—can be achieved since these so-called *anchor items* allow students' performances to be directly linked across years. This step is usually done in close collaboration between the head of each item-development group and the psychometric experts on the ÉpStan team.



	Level I		Level II		Level III		Level IV	
	Context A	Context B	Context A	Context B	Context A	Context B	Context A	Context B
Domain A	4	2	4	2	2	4	2	4
Domain B	4	2	4	2	2	4	2	4

Table 2. Exemplary test framework defining the minimum number of items per domain, context, and level to be included in the main test to ensure psychometric quality.

As soon as the final set of items is selected, the second step (2) of test compilation begins with the composition of the paper-pencil-based test booklets for grade 3 and the different school-form-related versions for grade 9. Again, certain content-related criteria are followed when deciding on the sequence of the items. For each test booklet/ version (a) relatively easy items are presented at the beginning to ensure motivation and to avoid early frustration among the students, and (b) item difficulty, response format, content, and assessed competency are alternated for the items that follow. After finalization of the tests, rigorous peer-review guarantees that these criteria are fulfilled. If content-based criteria cannot be fulfilled, an adaptation of the set of items chosen in step 1 is considered.

1.2.2 Specifics of Test Compilation

1.2.2.1 Grade 3

In grade 3, the ÉpStan tests are assessed by means of paper-pencil-based test booklets. As the coding of the students' answers is done by the teachers themselves, explicit coding instructions and coding booklets are developed together with the test booklets. Before going to print, the test booklets are designed and type-set by a professional illustrator. To ensure high quality, all documents have to pass through a multi-step proofreading process.

In mathematics, to address the higher number of different competencies and to obtain a more reliable picture of students' mathematical abilities, assessment of these competencies is done with two different test booklets (representing 50 minutes of testing time each) that are administered on two different days. The two test booklets together encompass approximately 70 items.

Compared to mathematics, German reading comprehension and German listening comprehension are assessed with one test booklet each on different days. Each test booklet encompasses approximately 40 items. Whereas the administration of the German reading comprehension test takes 50 minutes, the administration of the German listening comprehension test takes about 40 minutes. The administration of the latter is ensured by the accompanying CD, which includes not only the test content (stimuli and items) but also the test instructions themselves.

1.2.2.2 Grade 9

In grade 9, three different test versions are created. These three versions contain different proportions of easy, medium, and difficult items in order to adequately tackle the assumed competency level of different secondary school form populations. Hence, the EST-PREP version contains a higher number of level-1 items, whereas the ES version includes more level-4 items. The EST version, in comparison, includes a majority of level-2 and level-3 items. In addition, special attention is paid to items that serve as anchor items across the three test versions: These items (at least one third of all the items) are the same for two or three school forms and allow performances of students from the three school forms to be compared with each other. Within the sequence of all the items, these anchor items are presented in the same position on each test version to ensure the comparability of the yielded results. Note that grade 9 thus contains longitudinal anchor items (see Section 1.2.1) and anchor items across the three test versions.

The test administration in grade 9 is completely computer-based. In mathematics, students are allowed to choose the language of the test (German or French) and may also change the language throughout the test. Corresponding instructions regarding the test platform are given at the beginning of the test.

2 Data Processing, Analyses, and Reporting

Dalia Lorphelin, Ulrich Keller, Antoine Fischbach, Martin Brunner

2.1 Data Processing

2.1.1 Data Acquisition

2.1.1.1 Grades 3 and 7

In grades 3 and 7, data were collected using paper-and-pencil tests exclusively. In grade 3, students worked on test booklets, which teachers graded by marking the results on *coding sheets*. Additionally, students filled out *student questionnaires* and took home *parent questionnaires* to be filled out by their parents or guardians. In grade 7, students filled out only the student questionnaire.

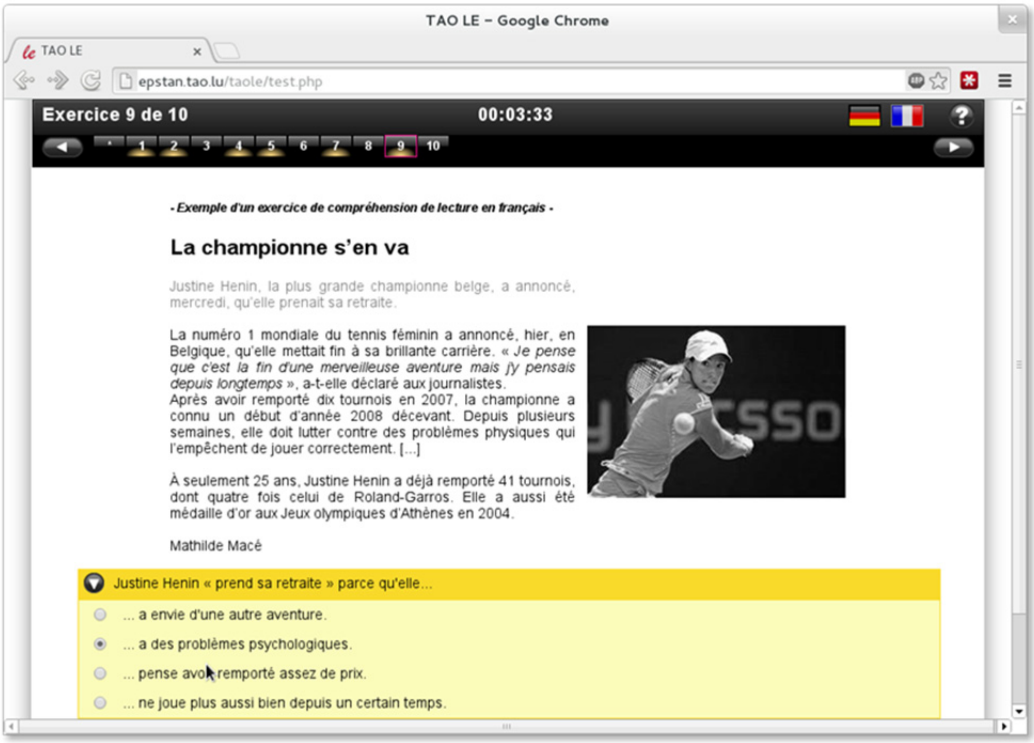
The completed coding sheets, student questionnaires, and parent questionnaires were then sent back to the ÉpStan team by the teachers. The sheets were then scanned, and the information they contained was retrieved using optical mark and character recognition (through the Teleform forms processing application).

2.1.1.2 Grade 9

All tests were administered on the schools' computers using the in-house *Online Assessment System* (OASYS; formerly labeled taoLE; Figure 1).

OASYS's principal design goals were robustness, prevention of data loss, ease of use, multilingualism, and visual attractiveness. To ensure robustness and data integrity, the client software running in a web browser immediately sends each response given by a student to the central server and awaits the server's confirmation that the response has been stored in the database. While the confirmation is pending, any additional responses given by the student are queued. If the confirmation times out, the client tries sending the response again. If this fails for a certain amount of time, the client assumes that it has been disconnected from the server and displays a message informing the student of the technical problem and providing directions to supervising school personnel about how to resolve the situation.

On the server side, the testing setup consists of two redundant web servers and two redundant database servers configured such that if one fails, the other will continue operating seamlessly.



- Exemple d'un exercice de compréhension de lecture en français -

La championne s'en va

Justine Henin, la plus grande championne belge, a annoncé, mercredi, qu'elle prenait sa retraite.

La numéro 1 mondiale du tennis féminin a annoncé, hier, en Belgique, qu'elle mettait fin à sa brillante carrière. « *Je pense que c'est la fin d'une merveilleuse aventure mais j'y pensais depuis longtemps* », a-t-elle déclaré aux journalistes.
Après avoir remporté dix tournois en 2007, la championne a connu un début d'année 2008 décevant. Depuis plusieurs semaines, elle doit lutter contre des problèmes physiques qui l'empêchent de jouer correctement. [...]

À seulement 25 ans, Justine Henin a déjà remporté 41 tournois, dont quatre fois celui de Roland-Garros. Elle a aussi été médaille d'or aux Jeux olympiques d'Athènes en 2004.

Mathilde Macé

Justine Henin « prend sa retraite » parce qu'elle...

○ ... a envie d'une autre aventure.

● ... a des problèmes psychologiques.

○ ... pense avoir remporté assez de prix.

○ ... ne joue plus aussi bien depuis un certain temps.

Figure 1. The testing client running in a web browser

2.1.2 Linking Student Data

In general, the links between students' test IDs and their test data were established in the following manner:

- The ÉpStan team issued *test IDs* that were not yet linked to individual students.
- The test IDs were linked to students' personally identifying information (*PII)* by the participating schools.
- Test IDs together with the PII were transmitted to the *Luxembourg Ministry of Education and Vocational Training* (MENFP).
- The MENFP matched *unique IDs* to the data received from schools. These were internal database IDs used in the MENFP's student databases.
- Data containing test IDs, unique IDs, and additional information from the student databases, but no PII, were transmitted to the ÉpStan team.

Since the unique IDs remain unchanged over the course of a student's school career (except for the transition from primary to secondary school), they can be used to build a pseudonymous longitudinal database that does not contain any PII.

The precise implementation of this scheme differed across grades:

- In grade 3, each class received a printed list with numeric test IDs. The same test IDs were printed on all test material (test booklets, coding sheets, and student and parent questionnaires). Teachers completed the list with the students' PII, handed out the test material according to the list and, after testing, sent the list to the MENFP, keeping a copy for themselves. The unique ID added by the MENFP was the ELE_ID key from the SCOLARIA primary school database.

- In grade 7, the printed questionnaires' first sheet contained a numeric test ID. Students filled in their PII and separated the first sheets from the rest of the questionnaires. The first sheets were sent to the MENFP, whereas the remaining sheets were sent to the ÉpStan team. The database key ELE_ID from the secondary school database *fichier élève* figured as the unique ID.

- In grade 9, schools were issued Excel files containing multiple worksheets, one for each class. Each worksheet contained a list of logins at the top, with empty columns for students' PII. Below, a series of login sheets automatically incorporated the information added using cell references. The schools' secretariats filled in the PII and printed out the login sheets, which were handed out to the students prior to testing. Each student thus received his or her own personal login, which was linked to his or her PII. The completed Excel files were sent to the MENFP via e-mail. As in grade 7, the ELE_ID from *fichier élève* was added as a unique ID.

2.1.3 Data Cleaning

2.1.3.1 Grades 3 and 7

After running optical mark and character recognition on the scanned questionnaires and coding sheets, the Teleform software required that an operator review and resolve all ambiguities (e.g., multiple tick marks where only one is expected). In addition, the data resulting from the coding sheets were checked for inconsistencies and the scanned documents were consulted to resolve any problems. A small amount of data that was missing because of a malfunction of the Teleform software was entered by hand.

2.1.3.2 Grade 9

All data were retrieved directly from the OASYS database server. No further data cleaning was required.

2.1.4 Scoring of Responses

All items were scored dichotomously, i.e., a response was considered to be either correct or incorrect.

In grade 3, no scoring of responses to competency test items was necessary as this had already been done by the teachers administering the tests. The reliability of teachers' scoring was verified by re-scoring a substantial random sample of test booklets, which revealed a high degree of consistency (K = 95.3).

In grade 9, multiple-choice items were scored by comparing the correct response to the one given by the students.

For short answer items, where the correct response was always a number or a fraction, students' responses were preprocessed before making this comparison by removing extraneous text such as white space, repetitions of the question, and units. Commas were replaced by dots (decimal separator). When the correct response was a fraction, students' responses were converted to floating point numbers and compared with the correct response in the same format. Comparisons were tested for "near equality" to account for the inherent lack of precision of floating point representations. Test authors verified that all unique responses were correctly classified as correct or incorrect.

2.1.5 Construction of Socio-Economic Indicators

2.1.5.1 Grade 3

Students were asked to take home a parent questionnaire that asked for parents' highest educational degree and occupation. Both questions were given in a multiple-choice format. For occupations, the choices were derived from the ISCO-88 classification and transformed into the ISEI-88 measure (see Ganzeboom, & Treiman, 1996). ISCO-08 could not be used as the necessary data were not yet available for Luxembourg.

2.1.5.2 Grades 7 and 9

Three socio-economic indicators were derived from students' responses in the student questionnaire:

- Wealth is a measure of information about the number of certain items (cars, bathrooms, etc.) available at the students' homes.
- Number of books is a single-item measure asking students to estimate the total number of books in their household excluding school books.
- ISEI-08 is a measure of occupational status derived from coding students' responses regarding their parents' occupations into the ISCO-08 classification (see Ganzeboom, 2010).

2.2 Test Scaling and Anchoring

Reporting the results of different tests on the same scale is essential for performing trend analyses (i.e., comparing pupils' outcomes across time). In the following, we will provide a detailed description of the procedure applied to scale and longitudinally anchor the annually collected ÉpStan data.

Each year, two cohorts were considered:

- Cohort 1: the ongoing year's data;
- Cohort 0: all the preceding years' data (beginning with 2010);

and a five-step procedure (see Nagy & Neumann, 2010) was followed:

- Step 1: Rasch compliance
- Step 2: Descriptive DIF analysis
- Step 3: ETS DIF analysis
- Step 4: Sensitivity analysis
- Step 5: Final estimation of person parameters

2.2.1 Step 1: Rasch Compliance

In this first step, we selected a set of Rasch-compliant items for each test. Cohort 1's data were scaled and items were freely calibrated (i.e., constraints were placed on cases; see Wu, Adams, Wilson, & Haldane, 2007). Only Rasch-compliant items (e.g., Bond & Fox, 2010; see also Gustafsson, 1980; Martin-Löf, 1974; Wright, Linacre, Gustafsson, & Martin-Löf, 1994) were kept in the models; that is

- items with a weighted MNSQ ≥ 0.8 *and* ≤ 1.2;
- items with a discrimination ≥ 0.25.

2.2.2 Step 2: Descriptive DIF Analysis

In Step 2, we graphically compared the item difficulties of potential anchor items—that is, items that were available in the data of both Cohort 1 and Cohort 0. To do so, the item difficulties for Cohort 1—as estimated in Step 1 (see Section 2.2.1)—were plotted against the item difficulties for Cohort 0 (Figure 2). A 95% confidence interval was computed for each item difficulty $(estimate \pm 1.96\, se)$. This allowed for a quick graphical evaluation of item robustness across cohorts and thus helped us to identify items that might show the problem of *differential item functioning* (DIF).

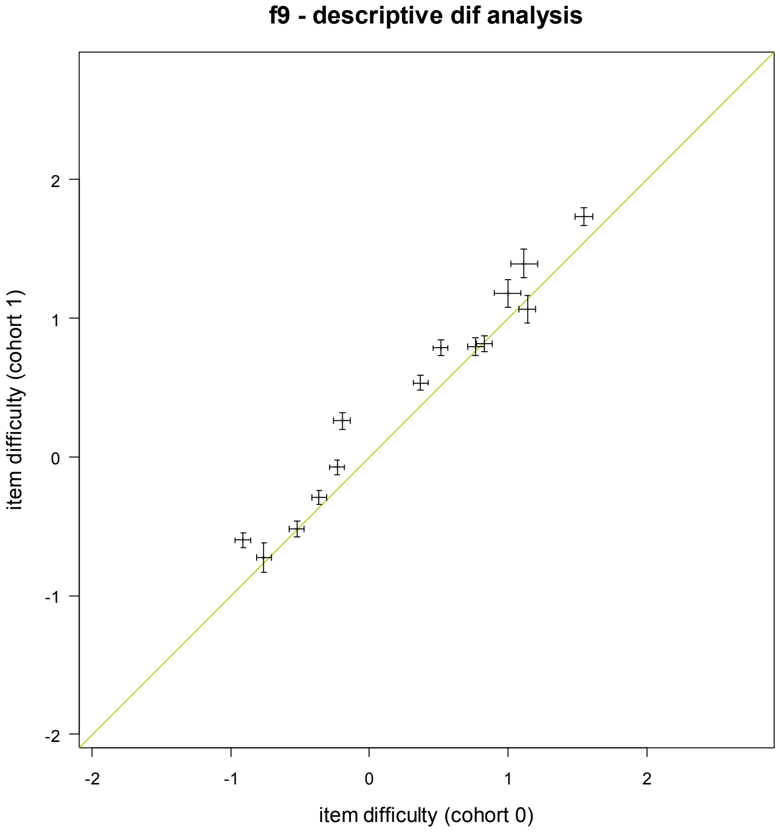


Figure 2. Descriptive DIF analysis exemplified for the 2012 Grade-9 French reading comprehension test, using the 2012 data as Cohort 1 and the pooled 2011 and 2010 data as Cohort 0.

For each test, a summary table (Table 3) regrouping all potential anchor items was produced.

code.1	id. content	id.irt	difficulty. 1	difficulty. 0	se.1	se.0
f905511c	3	3	1.732	1.546	0.034	0.033
f905521c	4	4	0.815	0.828	0.029	0.028
f905541c	5	5	1.065	1.139	0.051	0.030
f90c511c	41	42	-0.601	-0.909	0.028	0.029
f90c522c	42	43	0.784	0.515	0.029	0.027
f90c531c	43	44	0.534	0.369	0.028	0.027
f90c541c	44	45	-0.293	-0.361	0.027	0.027
f90c551c	45	46	-0.075	-0.233	0.027	0.027
f90c561c	46	47	-0.520	-0.522	0.028	0.027
f90c571c	47	48	-0.727	-0.760	0.054	0.028
f952121c	51	52	1.393	1.114	0.052	0.049
f952151c	52	53	1.181	0.998	0.051	0.049
f958141c	57	58	0.260	-0.196	0.031	0.030
f958151c	58	59	0.795	0.767	0.032	0.031

Table 3. Summary table exemplified for the 2012 Grade-9 French reading comprehension test. code.1 = item code. id.content & id.irt = database identifiers for code.1. difficulty.1 & difficulty.2 = item difficulties in Cohort 1 and Cohort 0, respectively. se.1 & se.0 = standard errors of difficulty.1 and difficulty.0 estimates, respectively.

2.2.3 Step 3: ETS DIF Analysis

In this third step, an anchored data table was built and scaled. The resulting person parameters (WLE scores; Warm, 1989) were used to investigate DIF by applying a logistic regression. For the sake of clarity, a typical construction design is provided in Table 4.

		2012 test				2011 test		2010 test
		Potential anchor items			Specific items	Anchor items	Specific items	Specific items
cases	2012	0/1	0/1	0/1	0/1	9	9	9
	2011	0/1	0/1	9	9	0/1	0/1	9
	2010	0/1	9	0/1	9	0/1	9	0/1

Table 4. Typical construction design of an anchored data table exemplified for the 2012 tests, using the 2012 data as Cohort 1 and the pooled 2011 and 2010 data as Cohort 0. 9 = nonadministrated items. 0 & 1 = incorrect and correct answers (see Section 2.1.4), respectively.

A cohort indicator variable and person parameter estimates (WLE scores) were used to explain the log odds of giving a correct answer to a given item. More formally:

∀ potential anchor item j, we have

$$logit\left(P(X_{ij}=1)\right)=\beta_{0j}+\beta_{1j}wle_i+\beta_{2j}cohort_i+e_{ij} \qquad \forall \text{ case } i$$

with

$$X_{ij}=\begin{cases}1\ if\ case\ i\ answers\ item\ j\ correctly\\ \qquad\qquad 0\ otherwise\end{cases}$$

$$cohort_i=\begin{cases}1\ if\ case\ i\in \text{cohort } 1\\ 0\ if\ case\ i\in \text{cohort } 0\end{cases}$$

$\beta_{0j},\beta_{1j},\beta_{2j}$ $are\ the\ regression\ coefficients\ indicating\ the\ relative\ effect\ of\ the$

$corresponding\ variable$

e_{ij} $is\ the\ error\ term$

The items were then classified according to the results of statistical hypothesis testing of the β_{2j} coefficient. Three disjointed sets were distinguished (see Nagy & Neumann, 2010):

- ETS Type A: DIF-free items,

  β_{2j} significantly different from 0.

- ETS Type B: items with moderate DIF,

  $|\beta_{2j}|$ significantly lower than 0.4.

- ETS Type C: items with large DIF,

  not in category A or B

Another finer item classification was defined on the basis of β_{2j}'s absolute value:

$$|\beta_{2j}|\le c\ where\ c\in\{\,0.1,0.2,0.3,0.4\,\}$$

The above-mentioned analysis results were assembled into recapitulation tables. An example is provided in Table 5.

Item code	beta2 estimate	std.error	p-value	ci.inf	ci.sup	ets.type	|beta2. lt.0.1|	|beta2. lt.0.2|	|beta2. lt.0.3|	|beta2. lt.0.4|
f905511c	-0.076	0.041	0.063	-0.156	0.004	1	1	1	1	1
f905521c	0.081	0.034	0.017	0.014	0.148	2	1	1	1	1
f905541c	0.023	0.059	0.699	-0.093	0.137	1	1	1	1	1
f90c511c	-0.129	0.046	0.005	-0.219	-0.039	2	0	1	1	1
f90c522c	-0.158	0.040	0.000	-0.237	-0.080	2	0	1	1	1
f90c531c	-0.046	0.040	0.242	-0.124	0.031	1	1	1	1	1
f90c541c	0.067	0.040	0.093	-0.011	0.145	1	1	1	1	1
f90c551c	-0.075	0.037	0.045	-0.148	-0.002	2	1	1	1	1
f90c561c	0.190	0.042	0.000	0.108	0.273	2	0	1	1	1

Table 5. DIF analysis from a logistic regression exemplified for the 2012 Grade-9 French reading comprehension test. *p*-value = test of significance of $\boldsymbol{\beta_{2j}}$, testing $\boldsymbol{H_0}: \boldsymbol{\beta_{2j} = 0}$ against $\boldsymbol{H_1}: \boldsymbol{\beta_{2j} \neq 0}$ at risk $\boldsymbol{\alpha = 0.05}$. ci.inf & c.sup = lower and upper confidence interval boundaries of $\boldsymbol{\beta_{2j}}$ at risk $\boldsymbol{\alpha = 0.05}$. ets.type 1 = ETS type A items. ets.type 2 = ETS type B items. ets.type 3 = ETS type C items.

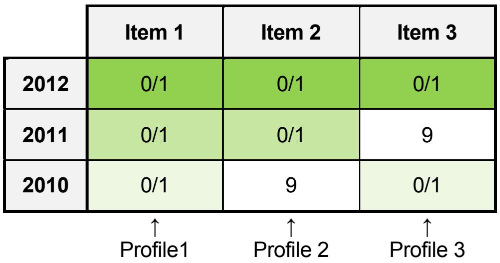
2.2.4 Step 4: Sensitivity Analysis

Several anchoring scenarios were defined depending on the retained potential anchor items. Each possible scenario lay within the following exhaustive set

$$S = \{ETS\ type\ A,\ ETS\ type\ A \cup ETS\ type\ B, all\ anchor\ items, |\beta_{2j}| \leq 0.1,\ |\beta_{2j}| \leq 0.2, |\beta_{2j}| \leq 0.3, |\beta_{2j}| \leq 0.4\}.$$

If a potential anchor item was discarded in a given scenario, it was included as a *virtual item* in the anchored data table (Table 4). For the sake of comprehension, we will illustrate this procedure with a practical example:

Suppose Cohort 1 and Cohort 0 represent the 2012 and the pooled 2011 and 2010 data, respectively. There are three possible profiles for any potential anchor item (Table 6).

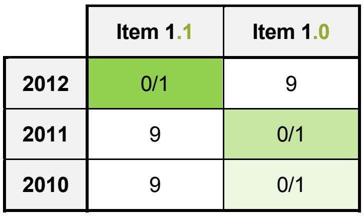


	Item 1	Item 2	Item 3
2012	0/1	0/1	0/1
2011	0/1	0/1	9
2010	0/1	9	0/1
	↑ Profile1	↑ Profile 2	↑ Profile 3

Table 6. Possible profiles for potential anchor items.

If Item 1—a profile-1 item—is not kept in the anchoring scenario, two possibilities can be distinguished:

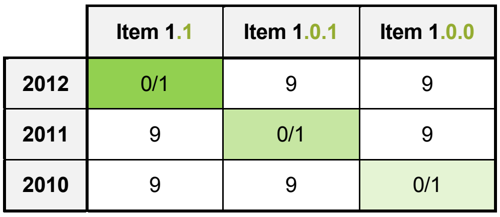
- Item 1 was used in the final model for anchoring the 2011 and 2010 data. If this is the case, Item 1 is split into 2 virtual items: Item 1.1 and Item 1.0 (Table 7), where Item 1.1 is freely estimated and Item 1.0 is constrained.



	Item 1.1	Item 1.0
2012	0/1	9
2011	9	0/1
2010	9	0/1

Table 7. Scenario 1 exemplified for a profile-1 item.

- Item 1 was *not* used in the final model for anchoring the 2011 and 2010 data. If this is the case, Item 1 is split into 3 virtual items: Item 1.1, Item 1.0.1, and Item 1.0.0 (Table 8), where Item 1.1 is freely estimated, and Item 1.0.1 and Item 1.0.0 are constrained.



	Item 1.1	Item 1.0.1	Item 1.0.0
2012	0/1	9	9
2011	9	0/1	9
2010	9	9	0/1

Table 8. Scenario 2 exemplified for a profile-1 item.

The same line of reasoning applies to profile-2 and profile-3 items (Table 6).

Following these rules, seven anchored data tables (Table 4) were built for each of the seven possible anchoring scenarios (Figure 3; see also Table 5) and scaled accordingly. At this stage, we needed to arbitrate and choose the optimal anchor item set to use. The rationale was to

- include a maximum number of anchor items, while
- ensuring that the mean differences remained robust—Cohen's *d* (1992) was calculated for each scenario—across the various invariance scenarios (Figure 3).

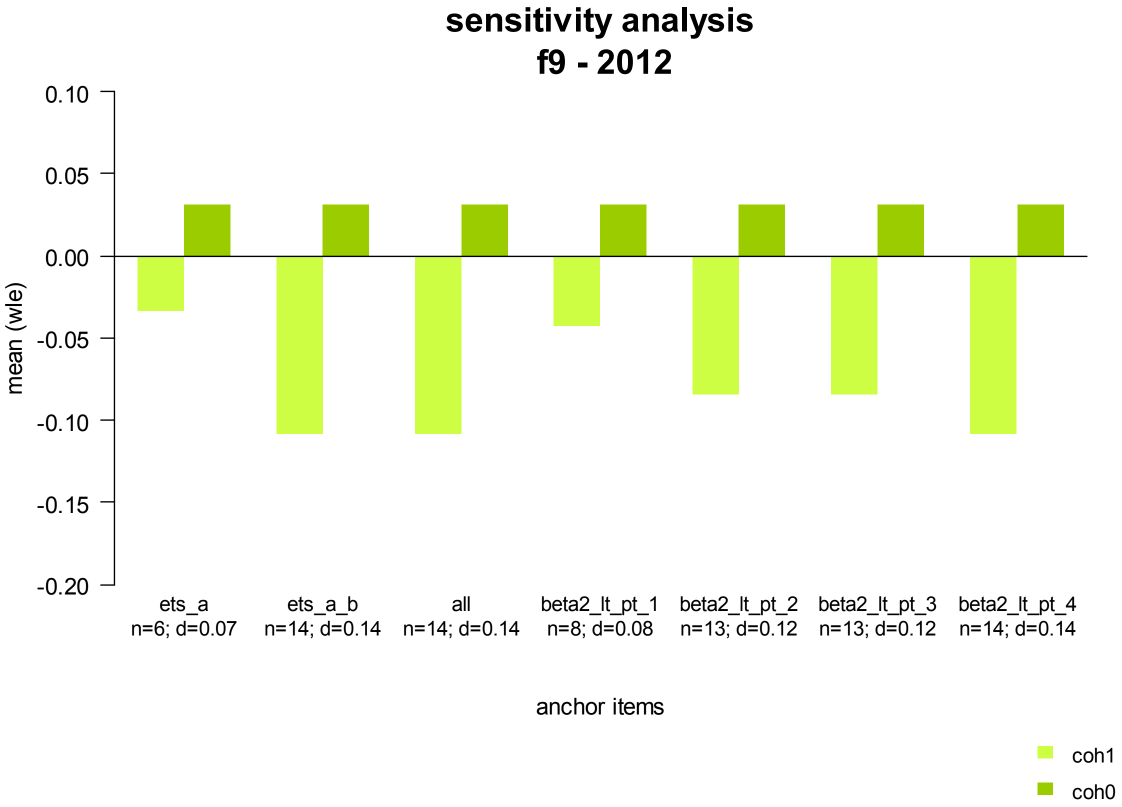


Figure 3. Sensitivity analysis exemplified for the 2012 Grade-9 French reading comprehension test.

2.2.5 Step 5: Final Estimation of Person Parameters

Once the optimal anchor item set was fixed, we proceeded to the final estimation of the person parameters by scaling Cohort 1's data (Table 9).

	Final anchor items	Remaining items (2012 test)
2012	0/1/9	0/1/9

Table 9. Final estimation of the person parameters exemplified for the 2012 tests.

Importantly, for this final estimation, the final anchor item parameters were constrained (i.e., imported from the optimal anchoring scenario run; see Section 2.2.4), while all remaining items were freely calibrated.

2.3 Calculation of Cut Scores for Proficiency Levels

For each test, the items were regrouped by theoretically defined and empirically validated—during the pretests—proficiency levels (as defined in Section 1). Based on the item difficulties (as estimated in Section 2.2), an outlier-free median difficulty[5] $\tilde{b}$ was inferred for each level. This median difficulty served as the basis for the calculation of the cut score CS_{level}.

If students are attributed a specific proficiency level, this implies that they have mastered—with high probability—the majority of the items on this level. In line with the *Programme for International Student Assessment* (PISA; OECD, 2009, p. 300), this "high probability" is operationalized as 0.62. Hence, students are attributed a high proficiency level if they have a 62% chance of correctly answering the virtual test question located at $\tilde{b}$. More formally, given the Rasch model, we have

$$0.62 = \frac{e^{CS_{level}-\tilde{b}}}{1+e^{CS_{level}-\tilde{b}}}$$

the cut score CS_{level} can be derived as

$$CS_{level} = log(0.62) - log(1-0.62) + \tilde{b}$$

where

$$\tilde{b}\ is\ the\ level's\ outlier\text{-}free\ median\ difficulty$$

On the basis of the final anchored competency estimates (see Section 2.2, and more specifically Section 2.2.5) and the cut scores (as defined above), each student is attributed a proficiency level, and the population's competency level distribution is inferred (Figure 4).[6]

As the item pools increase each year, the actual cut scores inevitably shift a bit from year to year.[7] However, this shifting entails that proficiency levels are then not truly comparable from one year to another, or alternatively that they must be readjusted backwards each year to ensure comparability across time. Neither of these two scenarios is appropriate for intelligibly communicating trends in student competencies, and this is the primary objective of proficiency levels. That is why the cut scores were fixed in the baseline cohort, that is, the pooled 2011 and 2010 cohorts.

[5] For each level, the item difficulties are box-plotted. Outliers, which are items that are at least 1.5 times the interquartile range above the third or below the second quartile, are not considered so that the best possible measure of a midpoint can be determined.

[6] Across all grade levels, test domains, and difficulty levels, the distance between two adjacent cut scores was 0.7 logits (range 0.4–1.1 logits) on average.

[7] Across all grade levels, test domains, and difficulty levels, this shift is currently an average of 0.1 logits (range 0.0–0.3 logits), with lower stability at the extremes of the difficulty distribution.

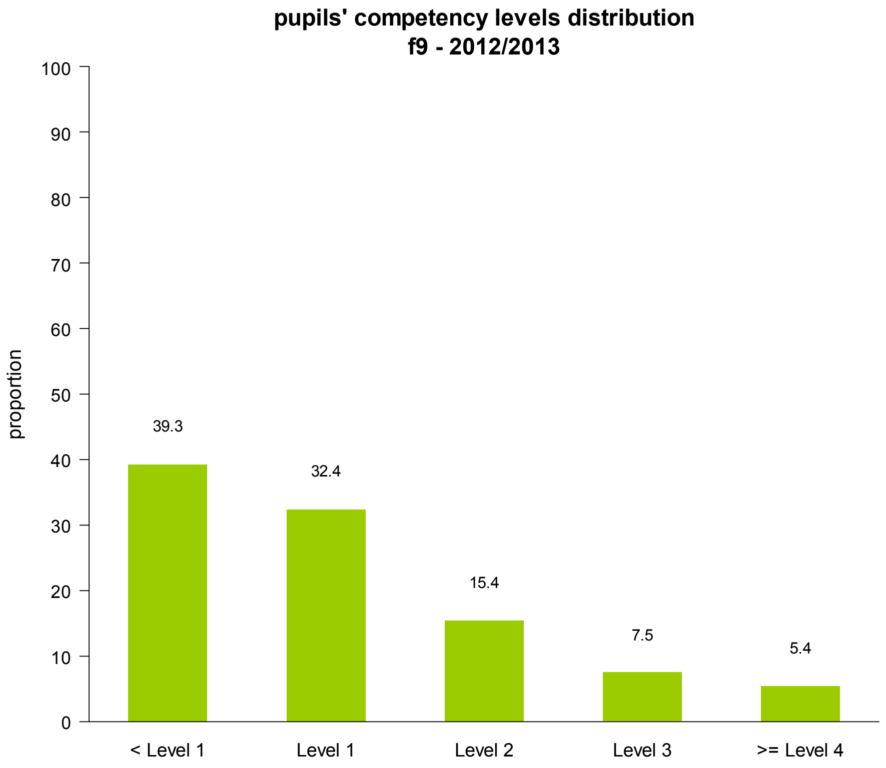


Figure 4. Proficiency levels exemplified for the 2012 Grade-9 French reading comprehension test.

2.4 Calculation of Expected Ranges (Fair Comparisons)

To enable educators to judge the performance of their school or class against groups of students with similar characteristics, we used student characteristics[8] (see Figure 6) to compute so-called *expected ranges* of performance at the school and class levels. The expected ranges were calculated in the following way (see also Robitzsch, 2011):

Using a data set containing data from all students who took at least one test, missing data were multiply imputed under a linear mixed model using the software package pan (Zhao & Schafer, 2012) for the statistics environment R (R Core Team, 2012). 30 imputed data sets were created.

These imputed data sets were aggregated at the class or school level, respectively. The aggregation was done separately for each testing domain, including only cases that had non-imputed data for the respective domain.

Linear regression models were fit to each data set, regressing the aggregated performance on the aggregated student characteristics.

For each data set, the predicted (fitted) values were computed for each case (school or class). The expected range is a 90% confidence interval around the predicted values (combined across data sets). The confidence intervals incorporated measurement error, the regression model's prediction error, and imputation error (error due to missing data).

The measurement error at the aggregate level for unit (school or class) j is given by

$$s_{1j} = \sqrt{\frac{1}{n_j^2} \sum_{p=1}^{n_j} SE\left(\Theta_{pj}\right)^2},$$

where n_j is the number of students with a non-imputed result in unit j, and $SE(\Theta_{pj})$ is the standard error associated with the WLE score of student p in unit j. This is combined with the prediction error for each imputed data set i, $SE(\gamma_{ij})$ to form the error of the expected result:

$$s_{2ij} = \sqrt{s_{1j}^2 + SE(\gamma_{ij})^2}$$

Combining across imputed data sets according to Schafer's (1997; see also Little & Rubin, 1987) method, the within-imputation variance is obtained as

[8] The student characteristics used are school form (implicitly), gender, languages spoken at home, immigration background, socio-economic background (HISEI, wealth index, number of books), birth year, and prior attendance at précoce, kindergarten, and primary school in Luxembourg.

$$\overline{U_j} = \tfrac{1}{m}\textstyle\sum_{i=1}^{m} SE\left(\gamma_{ij}\right)^2,$$

the between-imputation variance as

$$B_j = \frac{1}{m-1}\sum_{i=1}^{m}\left(\gamma_{ij} - \overline{\gamma_j}\right)^2$$

and the total variance as

$$T_j = \overline{U_j} + \left(1 + \tfrac{1}{m}\right) B_j.$$

The expected range can now be calculated as

$$\overline{\left(\gamma_j\right)} \mp 1.645 \cdot T_j.$$

2.5 Reporting the Results

2.5.1 Levels of Reporting

Reports were given at the four levels of the school system:

- National (system) level
- School level
- Class level
- Individual (student) level

The national report is published every three years and contains in-depth analyses, whereas results on the remaining levels are disseminated every year and contain the results of automated routine analyses.

2.5.2 Contents of School, Class, and Student Reports

At the school and class levels, the students' results were summarized in the following way:

- Number and percentage of students who reached each competency level
- Distribution of competency scores (on the ÉpStan metric[9]; Figure 5)
- Mean and expected range of school/class compared to the national mean (Figure 6)
- Mean scale scores for questionnaire scales

For the secondary-school-level reports, these summaries were presented separately by test version (i.e., ES, EST, and PR).

At the class level, the figures displaying competency scores and competency levels represent each student by a circle that contains this student's numeric ID (see Figure 5), which can be related back to the student's identity by his or her teacher via the class lists (see Section 2.1.2).

[9] To improve comprehension—and to avoid the reporting of potentially negative competency estimates—final WLE scores (see Section 2.2.5) were standardized to $M = 500$ and $SD = 100$ in the baseline year (i.e., 2010).

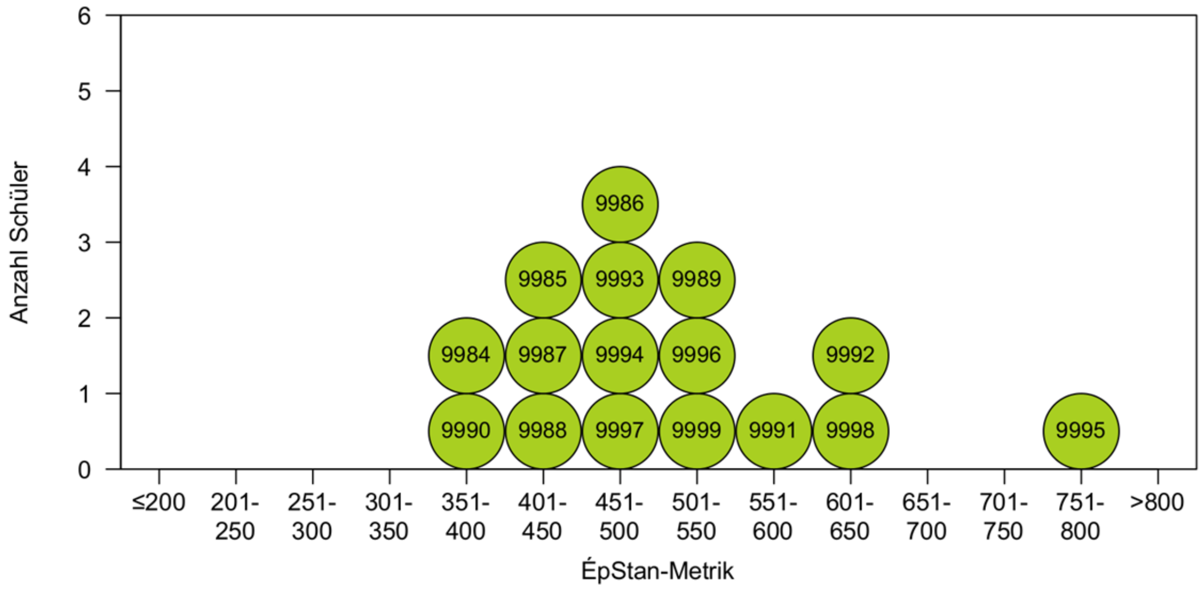


Figure 5: Example of a class-level plot of the competency score distribution. Circles represent individual students; numbers within circles are numeric student IDs. Anzahl Schüler = number of students. ÉpStan-Metrik = ÉpStan-metric.

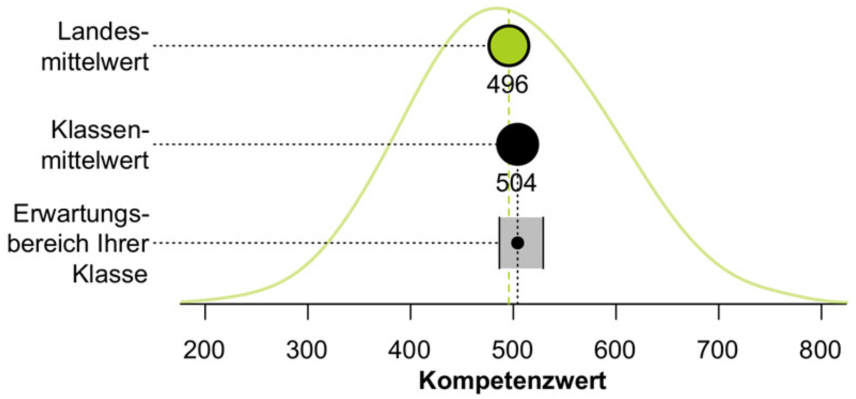


Figure 6: Example of a plot depicting a class mean, an expected range for the class, and the national mean. The curve in the background represents the distribution of scores for the whole sample. Landesmittelwert = national mean. Klassenmittelwert = class mean. Erwartungsbereich Ihrer Klasse = expected range for your class. Kompetenzwert = competency score.

Students received one single-page report per test domain depicting their competency score in relation to the national mean and the cut-offs that define the competency levels (Figure 7). In addition, the competency level the student attained was given in a short text along with descriptions of these competency levels.

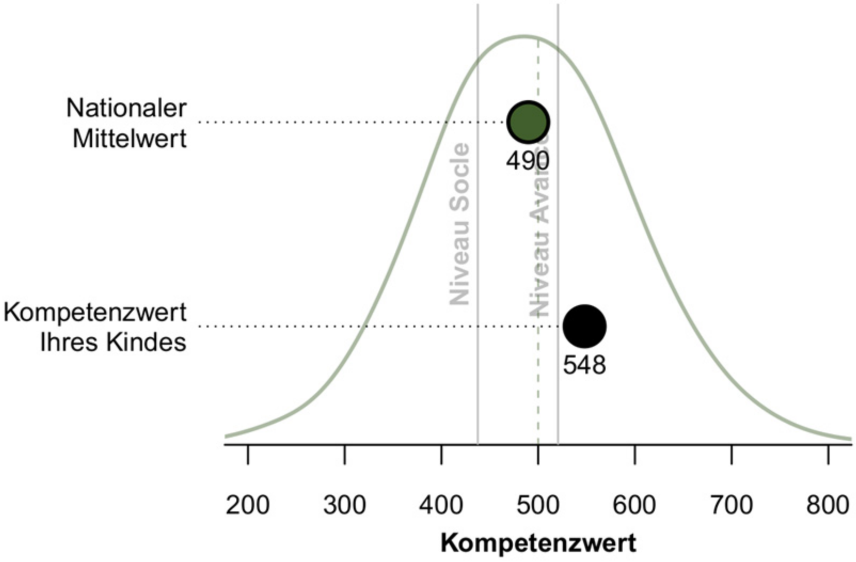


Figure 7: Example of a student-level plot depicting the student's competency score, the national mean, the competency level cut-offs, and the distribution of the competency scores for the total sample. Nationaler Mittelwert = national mean. Kompetenzwert Ihrer Kindes = your child's competency score. Niveau Socle = learning standard. Niveau Avancé = advanced learning standard. Kompetenzwert = competency score.

2.5.3 Dissemination of Reports

Hard copies of the national reports were distributed and were also freely downloadable from the ÉpStan and MENFP websites.

School-level reports were offered for download via the mySchool! online portal. All school presidents and inspectors (of primary schools) and principals (of secondary schools) were assigned unique logins to this platform through which they were offered the reports pertaining to their schools to download in a PDF format. For grade 9, hard copies of the school-level reports were distributed and were also e-mailed in PDF format upon request.

Class-level reports were offered for download via the mySchool! online portal. All teachers were assigned unique logins to this platform through which they were offered the reports pertaining to the classes and subjects they taught to download in a PDF format. This also included the student-level reports, which teachers were asked to print and hand out to their students.

In addition to the class and student reports, teachers could download documents containing detailed explanations and example items.

All documents contained a link to the ÉpStan website, which offers more in-depth information.

3 References

- Bond, T. G., & Fox, C. M. (2010). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York & London: Routledge.

- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037/0033-2909.112.1.155

- Ganzeboom, H. B. G. (2010). International Standard Classification of Occupations: ISCO-08 with ISEI-08 Scores. Retrieved from http://home.fsw.vu.nl/hbg.ganzeboom/isco08/isco08_with_isei.pdf

- Ganzeboom,, H. B. G., & Treiman, D. J. (1996). Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. *Social Science Research*, *25*, 201–239.

- Gustafsson, J.-E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *33*(2), 205–233. doi:10.1111/j.2044-8317.1980.tb00609.x

- Kühn, P. (2008). *Bildungsstandards Sprachen. Leitfaden für den kompetenzorientierten Sprachenunterricht an Luxemburger Schulen.* Luxembourg: MENFP.

- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

- Martin-Löf, P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics*, *1*(1), 3–18.

- MENFP. (2008a). *L'approche par compétences. Enseignement primaire. Cycle 2.* Luxembourg: MENFP.

- MENFP. (2008b). *Mathématiques. Division inférieure de l'enseignement secondaire. Compétences disciplinaires attendues à la fin de la classe de 6e et à la fin de la classe de 4e.* Luxembourg: MENFP.

- MENFP. (2010). *École fondamentale. Plan d'études. Édition 2010.* Luxembourg: MENFP.

- Nagy, G., & Neumann, M. (2010). Psychometrische Aspekte des Tests zu den voruniversitären Mathematikleistungen in TOSCA-2002 und TOSCA-2006:

Unterrichtsvalidität, Rasch-Homogenität und Messäquivalenz. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke, & K. Maaz (Eds.), *Schulleistungen von Abiturienten. Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand.* (pp. 281–306). Wiesbaden: VS.

- OECD. (2009). *PISA 2006. Technical report.* Paris: OECD Publishing.
- R Core Team. (2012). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.
- Robitzsch, A. (2011). *Technischer Anhang zur Rückmeldung der Baseline-Testung der 4. Schulstufe (2010).* Salzburg: BIFIE.
- Schafer, J. L. (1997). *Imputation of missing covariates under a multivariate linear mixed model* (Technical report No. 97-04). Dept. of Statistics, Pennsylvania State University.
- Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, *54*, 427–450.
- Wright, B. D., Linacre, J. M., Gustafsson, J.-E., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest. Version 2.0. Generalised item response modelling software.* Camberwell: ACER Press.
- Zhao, J. H., & Schafer, J. L. (2012). *pan: Multiple imputation for multivariate panel or clustered data.*