LE GOUVERNEMENT
DU GRAND-DUCHÉ DE LUXEMBOURG
Ministère de l'Éducation nationale,
de l'Enfance et de la Jeunesse

UNIVERSITY OF LUXEMBOURG
Luxembourg Centre for
Educational Testing (LUCET)

Explanations to the class and school report in the *Enseignement fondamental* and *secondaire*

1 Explanation to the competences

1.1 What is assessed by the ÉpStan? What is not?

The ÉpStan assess students' competences in key school areas using standardized tests. The tasks presented in these standardized tests are based on the educational standards developed by the MENJE for the different primary and secondary school cycles. In the educational standards, competences are defined as follows: *« la capacité de mettre en œuvre un ensemble organisé de connaissances, d'habiletés et d'attitudes qu'un élève doit mobiliser dans un contexte donné lui permettant de fournir une réponse adéquate à une problématique complexe »* (Jonnaert in MENFP, 2007, p. 3).

According to this definition, the best way to evaluate students' competences relies in presenting them with the task to solve complex problems. But what exactly does "complex" mean in this context? Solving complex problems requires the integrative use of knowledge and skills resulting in a preferably time-consuming and strenuous effort that calls for a high degree of motivation from the students.

Within the ÉpStan, we differentiate between different degrees of complexity depending on the requirements of the specific problem. The goal of a reading task can therefore be to localize and to name information explained in the text (lower complexity) or to draw a conclusion from what has been read (higher complexity).

It should however be emphasized that the ÉpStan competency tests do not assess all competences defined in the educational standards. On the one hand, not all competence areas can be assessed by the means of standardized tests. On the other hand, the time available to carry out the ÉpStan is very limited. Therefore, the ÉpStan are composed of a representative selection of key competences that are assessable in standardized ways and as economically as possible.

The ÉpStan tasks assessing reading and, if applicable, listening comprehension cover solely two key areas of the students' language abilities. These two key areas, however, play a central role in the students' everyday learning and school life insofar as receptive written language skills give access to further education. In addition, reading and listening comprehension can relatively economically be assessed in a highly standardized and objective manner.

Assessed competences in Cycle 2.1:

- **Mathematics in Cycle 2.1.** In terms of content, the mathematics tasks refer to the areas of (a) "space and shapes" (MENFP, 2011, p. 26, 116), (b) "numbers and operations" (MENFP, 2011, p. 28, 117), and (c) "sizes and measures" (MENFP, 2011, p. 30, 118).

- **Early literacy comprehension in Cycle 2.1.** The tasks evaluating early literacy comprehension (MENFP, 2011, p. 8) mainly refer to the following competences: phonological awareness, auditory attention, visual discrimination, and alphabetic knowledge (e.g., reading and writing letters or syllables).

- **Luxembourgish listening comprehension in Cycle 2.1.** The tasks evaluating listening comprehension in Luxembourgish (MENFP, 2011, p. 6) refer to various types of audio texts (e.g., dialogues, tales, reports). Tasks are developed for different competences (MENFP, 2011, p. 6), e.g., to understand one's interlocutor, to locate, understand and interpret information, or to apply listening strategies such as the recognition of noises.

- **German listening comprehension in Cycle 2.1.** In the German listening comprehension test in cycle 2 (MENFP, 2011, p. 6), students are confronted with various audio text formats (e.g. conversations, short stories or children radio shows). Test items and questions aim at various skills (cf. MENFP, 2011, p. 12): e.g. to follow a conversation and understand an interlocutor (who gives instructions), to understand short listening texts globally (as in this question: What are the children doing in the text?), to localize and understand individual pieces of information (as in the question: Where do the children meet?) and to employ listening strategies (e.g. by recognizing background noises).

Assessed competences in Cycle 3.1:

- **Mathematics in Cycle 3.1.** In terms of content, the mathematics tasks refer to the area of (a) "numbers and operations" (MENFP, 2011, p. 28-29, 121) and to a combined area of (b) "space and shapes" (MENFP, 2011, p. 26-27, 120) and "sizes and measures" (MENFP, 2011, p. 30-31, 122). In addition, the ÉpStan distinguish between two process competences assessing (a) "problem solving and modeling" (MENFP, 2011, p. 115) and (b) "specific basic skills" (MENFP, 2008, p. 11), which are defined as mathematical knowledge and skills that can be applied independently, without any context or transfer work. By differentiating between contextualized (problem solving and modeling) and decontextualized tasks (specific basic skills), the ÉpStan implicitly assess the content area of "solving arithmetic word problems" (MENFP, 2011, p. 32-33, 123).

- **German reading comprehension in Cycle 3.1.** The tasks evaluating German reading comprehension refer to two different types of texts: (a) continuous texts (e.g., stories, tales, factual texts) and (b) discontinuous texts (e.g., recipes, assembly instructions) are represented (similarly to PISA) in a two to one ratio. In test development, a further distinction is made between the two sub-competences of (a) "identifying and applying information" and (b) "construing information and activating reading strategies" (MENFP, 2011 p. 82-83).

- **German listening comprehension in Cycle 3.1.** The tasks evaluating German listening comprehension (MENFP, 2011, p. 12-13) refer to various types of audio texts (e.g., dialogues, reportages, radio plays). In test development, a distinction between the two sub-competences of (a) "identifying and applying information" (MENFP, 2011, p. 79-80) and (b) "construing information and activating listening strategies" (MENFP, 2011 p. 79-80) is being made.

Assessed competences in Cycle 4.1:

- **Mathematics in Cycle 4.1.** In terms of content, the mathematics tasks refer to the area of (a) "numbers and operations" (MENFP, 2011, p. 28-29, 125) and to a

combined area of (b) “space and shapes” (MENFP, 2011, p. 26-27, 124) and “sizes and measures“ (MENFP, 2011, p. 30-31, 126). In addition, the ÉpStan distinguish between two process competences assessing (a) “problem solving and modeling” (MENFP, 2011, p. 115) and (b) “specific basic skills“ (MENFP, 2008, p. 11), which are defined as mathematical knowledge and skills that can be applied independently, without any context or transfer work. By differentiating between contextualized (problem solving and modeling) and decontextualized tasks (specific basic skills), the ÉpStan implicitly assess the content area of “solving arithmetic word problems“ (MENFP, 2011, p. 32-33, 127).

- **German reading comprehension in Cycle 4.1.** The tasks evaluating German reading comprehension (MENFP, 2011, p. 16-17) refer to two different types of texts: (a) continuous texts (e.g., stories, tales, factual texts) and (b) discontinuous texts (e.g., reports, comments). They are represented (similarly to PISA) in a two to one ratio. In test development, a further distinction is made between the two sub-competences of (a) “identifying and applying information” and (b) “construing information and activating reading strategies” (MENFP, 2011 p. 88-89).

- **French reading comprehension in Cycle 4.1.** The tasks evaluating French reading comprehension (MENFP, 2011, p. 24-25) refer to two different types of texts: (a) continuous texts (e.g., stories, tales) and (b) discontinuous texts (e.g., recipes, advertisements). They are represented (similarly to PISA) in a two to one ratio. In test development, a further distinction is made between the two sub-competences of (a) “identifying and applying information” (MENFP, 2011, p. 104-105) and (b) “construing information and activating reading strategies” (MENFP, 2011 p. 104-105).

Assessed competences in 7e:

- **Mathematics in 7e.** The mathematics test in 7e covers the expected competences at the end of cycle 4 in the four sub-areas “numbers and operations” (MENFP, 2011, p. 28-29, 129), “space and shapes” (MENFP, 2011, p. 26-27, 128), „sizes and measures“ (MENFP, 2011, p. 30-31, 130) and “solving arithmetic word problems“ (MENFP, 2011, p. 32-33, 131) ab. The test consists of 23 to 28 short word problems, which are presented in French and German (students can switch languages at any time and thereby select the language they consider most appropriate to solve the problem). Most tasks are semi-open or multiple-choice. They are assumed to show to what extent students can correctly apply the taught knowledge in familiar and in new mathematical situations.

- **German reading comprehension in 7e.** The test evaluating German reading comprehension in 7e includes the two sub-competences “identifying and applying information presented in a text (Competence 1 – K1) as well as “construing information and activating reading strategies/techniques” (Competence 2 – K2; MENFP, 2011, p. 16-17). The test consists of four to six continuous (e.g., literary, or factual texts) or discontinuous (e.g., tables, illustrations, or maps) texts, each of an average length of 100 to 550 words and covering topics from students’ everyday life such as hobbies, society, friendship, or family. The test’s difficulty is based on the educational standards of the previous learning cycle in primary school (Cycle 4).

- **French reading comprehension in 7e.** The test evaluating French reading comprehension in 7e includes the two sub-competences "identifying and applying information presented in a text (Competence 1 – K1) as well as "construing information and activating reading strategies/techniques" (Competence 2 – K2; see competency grid and development stages, primary school, Cycle 1 to 4, MENFP, 2011, p. 22-23). The test consists of five to seven continuous (e.g., simple stories, tales, and factual texts) or discontinuous (e.g., illustrations, advertisements) texts, each of an average length of 50 to 500 words and covering topics from students' everyday life such as family, school, or friendship. The test's difficulty is based on the educational standards of the previous learning cycle in primary school (Cycle 4).

The duration of the test is 45 minutes per school subject for all students, with a varying test order.

Assessed competences in 5e:

- **Mathematics in 5e.** The test evaluating mathematical competences is focusing on tasks to assess mathematical models and problems, which can differ significantly in their level of complexity. All the presented tasks can be allocated to one of the following three mathematical content areas: (1) numbers and operations (MENFP 2008b; 2008c, p. 21-23), (2) figures of plane and space (MENFP 2008b; 2008c, p. 17-19), or (3) dependence and variation (MENFP 2008b; 2008c, p. 25-27). The students can solve the mathematical tasks in either German or French.

  German reading comprehension in 5e. The test evaluating German reading comprehension in 5e (MENFP 2008a, p. 34-35, 39-41, 50-52, 68) includes the two sub-competences "identifying and understanding information presented in a text" (Competence 1 – K1) as well as "analyzing and interpreting texts, and drawing (knowledge-based) conclusions" (Competence 2 – K2). These sub-competences can be derived from the description of the four learning levels in the educational standards (MENFP 2008a, p. 68). Each German test consists of five to eight continuous (e.g., stories, or factual texts) or discontinuous (e.g., tables, illustrations, or assembly instructions) texts of an average length of 100 to 750 words, covering topics such as hobbies, society, friendship, or vocational training.

  French reading comprehension in 5e. The test evaluating French reading comprehension in 5e also includes the two sub-competences "identifying and understanding information presented in a text" (Competence 1 – K1) as well as "analyzing and interpreting texts, and drawing (knowledge-based) conclusions" (Competence 2 – K2). Also for French, the sub-competences can be derived from the description of the four learning levels in the educational standards (MENFP 2008a, p. 68). Each French test consists of four to seven continuous (e.g., stories, or factual texts) or discontinuous (e.g., tables, or illustrations) texts of an approximate length of 100 to 750 words, covering topics such as friendship, hobbies, family, or vocational training.

Different test versions are created for the three school tracks "*Enseignement secondaire classique*", "*Enseignement secondaire général - voie d'orientation*" and "*Enseignement secondaire général - voie de préparation*", which contain a minimum of overlapping texts and exercises to ensure their comparability.

Accordingly, three test versions are created for each of the three areas of competence. The test versions are named in conformity with the respective school tracks:

- *Enseignement secondaire classique* (ESC)
- *Enseignement secondaire général – voie d'orientation (ESG)*
- *Enseignement secondaire général – voie de préparation (ESG-VP)*

On the one hand, each test version consists of tasks that are specifically tailored to the average performance level of the respective school track. On the other hand, each version furthermore includes a minimum number of tasks that all students can work on. By these means, comparisons of competencies across school tracks can be made with the help of modern psychometric methods.

The test time for all students is of 45 minutes per subject, irrespective of the test version, and the order of the tests varies randomly.

The ÉpStan thereby provide a highly standardized yet incomplete snapshot, whereas the *Bilans* or school grades are based on a long-term and complete evaluation of academic achievement. In this context, the results of the ÉpStan are to be seen as complementary to and not in competition with the *Bilans* or school grades.

1.2 How are competences measured?

In a first step, a theoretical level of difficulty is assigned to each task (e.g., "below *Niveau Socle*", "*Niveau Socle*", "*Niveau Avancé*", "above *Niveau Avancé*"). This initial classification takes place in the test development groups under a strict consideration of the respective descriptors (MENFP, 2011 & MENFP 2008a; 2008b; 2008c) and already before empirical indicators for the task difficulty are available. In a second step taking place after the data collection, the theoretical and empirical task difficulties are compared with each other. If these do not match for a specific task, either a theoretical reclassification is carried out - in case this seems justifiable based on the theoretical descriptors for the competence areas and levels in question – or the said task is not taken into consideration in the final calculation of the competence level.

Through this procedure that is described here in a simplified way, the test development groups can ensure that theoretically difficult or simple tasks are also empirically difficult or simple, respectively. In addition, it ensures that the empirical difficulties of the individual *Socles* or competence levels are relatively similar, so that, for example, all tasks used to calculate e.g. the *Niveau Socle* (in primary education) or e.g. the competence level 2 (in secondary education) in mathematics are roughly in the same empirical difficulty range.

1.3 What distinguishes the ÉpStan from conventional exams?

Conventional exams are primarily assessing knowledge taught in class that students should have mastered. This examination format assumes (in principle) that students who have attended the class should be able to solve all the tasks. If all tasks are correctly solved, the students fulfill the basic expectations – the target state. Conventional exams, however, do generally not go beyond this target state, so that they hardly offer very good

students any opportunities to prove their knowledge and skills. Such a conventional exam format only displays whether the students' performance fails to reach the desired target stage. This deficit-oriented approach is firmly anchored in our examination culture. When correcting the exams, little attention is paid to what students are doing right. Corrections are mainly highlighting what they are not (yet) succeeding in. In general, a conventional exam is thus rather aiming at identifying students' weaknesses or their deviation from the target state and less so at assessing what students actually can do.

In contrast, the ÉpStan offer a complete range of tasks with varying difficulty – covering what almost everyone can do, what everyone should (at a minimum) be able to do, and what goes (far) beyond that. It can thus not be expected from the outset that all students will be able to solve all ÉpStan tasks correctly. The students should rather make full use of their skills by working on as many tasks (of varying difficulty) as possible. The ÉpStan are thereby not aiming at identifying weaknesses, but at determining which (partial) sub-competences students already have.

In this respect, the examination approach pursued by the ÉpStan is clearly based on the principles of competence-oriented teaching. This does however not mean that the tasks used can generally be considered as representative for a competence-oriented form of teaching. The given circumstances are a little more complex.

The ÉpStan are aiming at a comparability of student performances with each other and with a given standard. Against this background, the assessed competence areas were selected because a) they are of central importance for everyday school and learning life and because b) they can be assessed in a standardized, objective, and reliable way with relatively little effort.

Against this background, the tested competence areas were selected because a) they are of central importance for everyday school and learning life and because b) they can be recorded in a standardized, objective, and reliable way with relatively little effort. The same hold true for the selected task formats. Therefore, the ÉpStan are mainly proposing closed and semi-open tasks. In the classroom setting, on the other hand, the focus is on the learning process of the students, and this should preferably be fostered through open task formats.

2 How reliable are the ÉpStan measurements? Study results from the *Enseignement fondamental*

A central aim of the ÉpStan is to record the students' competences as standardized as possible – i.e., objectively, and reliably – while taking the given framework into account. Measures, such as an exclusive use of closed and semi-open task formats, the provision of detailed instructions for correction (to avoid a too mild or too strict correction), and of a digital coding mask, are established to ensure that this central aim is achieved. In this context, the question of whether different correctors come to the same assessment of competences for individual students naturally arises.

To investigate this question, an additional study has been carried out in the scope of the ÉpStan 2010/11 (for 3.1) and of the ÉpStan 2015/16 (for 2.1 and 3.1). A random sample of 30 filled student booklets was drawn from each of the four test booklet versions. In a first step, all corrections made by the classroom teacher were made unrecognizable. In

a second step, a trained corrector who was strictly following the correction instructions given in the manual, corrected each of the 120 student booklets anew. As a result, assessments made by two correctors were available for 3400 (ÉpStan 2010/11), 3000 (ÉpStan 2015/16, Cycle 2.1) and 4400 (ÉpStan 2015/16, Cycle 3.1) student answers. The most important finding was that the two corrections were consistent to a very high degree and that this consistency remained stable over the two cohorts. The absolute level of agreement was at 97,4% (ÉpStan 2010/11), and at 97,2 (ÉpStan 2015/16, Cycle 2.1) and 98,0% (ÉpStan 2015/16, Cycle 3.1), respectively. Taking into consideration that the agreement between the correctors could also be a coincidence, the corrected level of agreement was still at 95.3% (ÉpStan 2010/11), and at 96,1 (ÉpStan 2015/16, Cycle 2.1) and 94,9% (ÉpStan 2015/16, Cycle 3.1), respectively. In summary, it can thus be concluded that the precise correction instructions and, above all, their careful application by teachers allow to record student competences in an extremely reliably and objective manner. Considering that the ÉpStan procedures from Cycle 3.1 are adopted almost one-to-one in Cycle 4.1, it can be assumed with a high degree of probability that the positive finding in Cycle 3.1 also applies to Cycle 4.1. In the *Enseignement Secondaire*, the students' answers are directly stored and processed on the test platform. A manual coding by teachers is thus not necessary.

3 Evaluation of the test results

3.1 ÉpStan-Metric

The overarching aim of the ÉpStan is to compare students' performance longitudinally – which means over time – in order to make potential effects of reforms in the educational sector visible. To achieve this goal, the same tests cannot be used in every school year, as otherwise there would be a risk that students would be specifically prepared for the test and that the test results would be distorted. Therefore, students' performance in the ÉpStan is not simply summed up into point totals, but competence values are calculated using statistical models from the so-called *Item Response Theory* (IRT). The competence values can in turn be compared across different cohorts and, if applicable, school tracks even though the competence tests entail different tasks each year. The only precondition for these comparisons is that some tasks have already been used in earlier test versions.

The competence values have been normed in such a way that the mean value for all the students in Luxembourg lies at 500 points (across tracks). The standard deviation (mean deviation of the test values from the mean value) was set at 100 points. Depending on the learning cycle, another reference year has been used, which is dependent on when the respective learning cycle has started to participate in the ÉpStan: For Cycle 2.1, the reference is the school year of 2014/2015, for Cycle 3.1 the year of 2011/2012, for Cycle 4.1 the year of 2017/2018 and in secondary school the reference year for 7e is 2018/2019[1] and for 5e 2011/2012, respectively. The indicated school year therefore serves as starting point and first reference year up to which this year's ÉpStan results can be compared. If the mean ÉpStan value of a specific student group changes from

[1] It has however to be considered that only approximately two-thirds of all 7e students have participated in the ÉpStan. When interpreting the results, it therefore needs to be taken into account that the assessed students do not represent the full 7e cohort and that the cut-off level for the determination of the *Niveau Socle* might be subject to change in the coming years.

one year to the next, it can be considered an indication that there is an actual change in the skills of that group.

3.2 What is the expectancy range?

Many teachers know in advance that the performance of their class will almost certainly be below or above the national average due to the classroom's socio-cultural and socio-economic composition. Hence, what is the point of assessing the performance if the results are seemingly determined beforehand.

Numerous international and national studies (e.g., PISA, PIRLS, the ÉpStan) have shown that characteristics such as gender, social background, and mother tongue clearly relate to performance in competence tests. In Luxembourg, a higher average performance in German reading comprehension is, for example, achieved by students who have Luxembourgish or German as a mother tongue, who have a high social status and/or who are female. On the other hand, students whose mother tongue is Portuguese or French, who have a low social status and/or who are male are on average performing lower in the competence area of German reading comprehension. The socio-cultural and socio-economic composition of the student population in Luxembourg differs however significantly between schools and classes, so that the performance of single classrooms cannot be easily compared.

The following (strongly simplified) example is explaining this situation in more detail: Let's imagine a fictitious school class that has participated in the German test. The students have achieved an average score of 430 points in the test (the points refer to the ÉpStan-metric described in section 3.1) and are thereby well below the national average. Let's furthermore assume that the class is mainly attended by male students from socially disadvantaged backgrounds who are speaking Portuguese or French at home. Against this background, the below-average performance of the class is not surprising. The more interesting question is therefore the following: How does the actual outcome of the class compare to the result that would have been expected given its student composition? To answer this question, we calculate "expectancy ranges" using a statistical model (the so-called regression model).

We do this as follows: In a first step, we determine the relation between the students' test performance in the measured competence areas on the one hand and the characteristics of gender, socio-economic background, mother tongue, migration background, school track (in secondary education) and school trajectory data on the other hand. We refer to these characteristics as background variables. For each of these background variables, the regression model provides a weight, a so-called parameter, that indicates how much that specific characteristic affects test performance.

In a second step, we use these parameters to calculate an expectancy range for every class, in which the result of the classroom should lie, provided that the relations depicted in the regression model reflect the actual composition of Luxembourg's student population. The width of this expectancy range reflects the measurement accuracy of the tests used, the prediction accuracy of the regression model, and the uncertainty due to missing data. If a class achieves a result above the expectancy range, it has thus performed better than would have been statistically expected based on the socio-economic and socio-cultural characteristics of the students. A result below the expectancy range, however, signalizes a worse result than would have been statistically expected.

In the scope of the ÉpStan, the performance profile of each individual class is thus related to classes with comparable socio-cultural and socio-economic student compositions. This approach allows, for example, to give teachers who work very effectively under “difficult” conditions positive and scientifically founded feedback on the performance level of their class. In principle, expectancy ranges are, however, of interest for all teachers who want to get a fair picture of their students’ competences.

4 Explanations on learning motivation

In addition to the performance tests, students are answering a questionnaire assessing the following areas:

- Learning motivation in the competence areas assessed in the ÉpStan
- General learning motivation in school
- Class and school climate

Students‘ learning motivation is assessed by statements such as “I am interested in most school subjects”. The students indicate their level of agreement with these statements by ticking a box. For this purpose, a four-point answer scale with answer categories ranging from "does not apply" to "applies" is available. The answers to the various statements are then thematically summed up into scale values (e.g., in a scale for general academic interest). In secondary school, students can complete the questionnaire in either German or French.

4.1 Learning motivation in school subjects

Learning motivation in the specific school subjects is assessed via four scales: academic self-concept, academic interest, school anxiety and reading interest (with the latter only for language subjects).

- The academic self-concept scale provides information on the students’ evaluation of their own academic performance and abilities in the respective school subject. To assess subject-specific academic self-concept, the students answer statements such as: “I am good at mathematics.”
- The academic interest scale indicates to what extent the students are interested in and enjoy a specific subject. An example item sounds as follows: “I am interested in mathematics.”
- The school anxiety scale provides information about the extent to which students are afraid of a subject, how much it worries them or in how far exams in the subject are making them nervous. An example item sounds as follows: “I am afraid of mathematics.”
- The reading interest scale indicates the extent to which the students like to read. An example item sounds as follows: “I like to read in German.”

Mean values are to be interpreted as follows: higher mean values represent a more positive academic self-concept, a higher academic interest in the subject or a higher reading interest - but also more anxiety in a certain school subject.

4.2 General academic learning motivation

The students' general academic learning motivation is assessed via three scales:

- General academic self-concept (e.g., „I am good at most school subjects.“)
- General academic interest (e.g., „I enjoy most school subjects.“)
- General school anxiety (e.g., „I am afraid of most school subjects.“)

Higher mean values indicate a more positive academic self-concept, a higher academic interest, and a stronger school anxiety.

4.3 Class and school climate

How students perceive their school, or their class is assessed via three scales.

- Teacher-student-relationship (e.g., „In my class, I get extra support from my teachers when I need it.“)
- Class climate (e.g., „In my class, we get along well.“)
- Tendency for disruptions (e.g., „In my class, we sometimes disrupt the class on purpose.“)

Higher mean values indicate a more positive teacher-student-relationship, a better class climate, but also a higher tendency for disruptions.

5 Referenzen

- MENFP (2007). *Français. Socle de compétences. Niveau fin 6e ES*. Luxembourg: MENFP.
- MENFP (2008). *L'approche par compétences. Enseignement primaire. Cycle 2*. Luxembourg: MENFP.
- MENFP (2011). *Plan d'études. École fondamentale*. Luxembourg: MENFP.
- MENFP (2008a). *Bildungsstandards Sprachen*. Luxemburg: MENFP.
- MENFP (2008b). *Mathématiques. Division inférieure de l'enseignement secondaire. Compétences disciplinaires*. Luxembourg: MENFP.

- MENFP (2008c). *Mathématiques. Division inférieure de l'enseignement secondaire technique. Compétences disciplinaires*. Luxembourg: MENFP.