Overall, a Good Test, but . . . —Swedish Lower Secondary Teachers’ Perceptions and Use of National Test Results of English

Gudrun Erickson* and Jörgen Tholin

Department of Education and Special Education, Faculty of Education, University of Gothenburg, 405 30 Gothenburg, Sweden; jorgen.tholin@gu.se
* Correspondence: gudrun.erickson@ped.gu.se

Abstract: This article builds on a study set within the Swedish educational system and focuses on lower secondary teachers’ use of national test results when awarding final grades of English as a foreign language (EFL). In Sweden, teachers are entrusted with the responsibility of assessing their own students’ competences as well as assigning grades. To support them, there are compulsory national tests to be used as important advisory tools; however, they are not exams in a strict decisive sense. After a brief contextualization and conceptualization regarding language education in Sweden, including the assessment, teachers’ somewhat contradictory perceptions and use of results from the national EFL test for 11–12-year-olds are described and discussed. Data emanate from large-scale teacher questionnaires conducted for three years (2013, 2016 and 2019), which are analyzed from quantitative as well as qualitative angles. Results indicate that a number of teachers struggle with factors related to the language construct as well as to the educational context and consequences at individual, pedagogical and structural levels. This is discussed from various angles, linked not least to the policy, curriculum and other frame factors. Furthermore, the need for further research in direct collaboration with teachers is emphasized.

Keywords: language education; Sweden; national assessment; EFL; young learners; teachers’ use of test results; final grades; validity; curriculum

1. Introduction

Since the publication of a revised curriculum for Swedish compulsory school in 1980, the national syllabuses for foreign language education (revised in 1994, 2000 and 2011) have been increasingly characterized by a communicative and functional approach (Canale and Swain 1980; Hymes 1972; Malmberg 2001). De facto, this has meant that language use and confidence have become focused upon, rather than single elements of language, for example, vocabulary, grammar and pronunciation, which are seen as important prerequisites rather than goals per se. Further, the connection to the Common European Framework of Reference for Languages (Council of Europe 2001) has gradually strengthened, although no formal alignment has been determined between the CEFR and the Swedish language syllabuses (Erickson and Pakula 2017).

The implementation of an action-oriented view of language competence has been conducted in different ways, in various information materials from the national educational authorities, through pre- and in-service teacher education and by means of teaching and study materials adapted to a functional view of language. In addition, the national assessment system (Section 1.3.2), which is intended to illustrate and measure the type and levels of competence required in the national curriculum, has been seen, at least implicitly, as a way to clarify and operationalize the syllabuses (Erickson and Åberg-Bengtsson 2012).

The current article focuses on English as a foreign language (EFL) in school Year 6, that is, for 11 to 12-year-old students, and reports on a study based on the compulsory
national test for this age group. The large majority of students and teachers were generally positive to these tests. However, a number of teachers seemed to make a consistently more severe interpretation of the proficiency levels required in the syllabus, not least regarding written production. This is manifested in their use of the national test results when awarding students’ final EFL grades in Year 6, but also in their comments on the test. Looking into and trying to better understand the underlying reasons for this, by analyzing teachers’ attitudes and written comments as demonstrated in the teacher questionnaire accompanying the test, is the main aim of this study. In addition, based on national statistics, teachers’ use of the national test results for grading purposes are reported on and briefly discussed.

1.1. Research Questions

Based on perceptions expressed in teacher questionnaires for three years of the Year 6 national test for EFL (2013, 2016 and 2019), the following research questions are addressed:

**Research Question 1 (RQ1).** What are Year 6 teachers’ responses to fixed-answer questions about the national tests of English?

**Research Question 2 (RQ2).** What are Year 6 teachers’ open comments on the national tests of English, focusing on the test as a whole and on written production?

**Research Question 3 (RQ3).** What indications regarding the use of the national test results for the final grading of English can be traced in the teachers’ responses and comments?

1.2. Conceptual Background and Previous Studies

The current study deals with language assessment, in particular, teachers’ perceptions and handling of test results. It is based, on the one hand, on theories of communicative competence and functionality (Canale and Swain 1980; Council of Europe 2001, 2020; Hymes 1972), and, on the other hand, on an expanded view of validity. This means that the use and consequences of assessment results are in focus, and it also implies a strong emphasis on values and ethics (Fox 2004; Kunnan 2004; Messick 1989). In addition, the frame factor theory (Lundgren 1999) is of importance, in particular focusing on different aspects of the curriculum. In this, van den Akker’s (2003) distinction between three broad categories—the intended, implemented and attained curriculum—was useful in interpreting and discussing observations and conclusions, as have reflections and research on pedagogical content knowledge and teacher cognition (Borg 2003; Shulman 1986).

Research on English in Swedish lower secondary schools, focusing on assessment and grading, is scarce. However, some studies conducted by the national educational authorities—the National Agency for Education/NAE (Skolverket) and the Swedish Schools Inspectorate (Skolinspektionen)—have been performed. A recent report from the Schools Inspectorate, focusing on the quality of teachers’ grading practices in school Year 6, and based on interviews with teachers, students and head teachers at 30 schools, concluded that teachers strive to gather and document comprehensive evidence for their grading, but that collaboration and co-rating ought to increase to enhance construct coverage and equity. It also concluded that the national test is taken into account only to a minor extent in grading. The actual relationship between national test results and grading is mentioned only in passing (Skolinspektionen 2020).

1.3. Contextual Background

Swedish compulsory school comprises 1+9 years, the first being a preschool class, normally starting at the age of six. The current national curriculum for compulsory school was introduced in 2011, and includes regulations for the preschool class, for Years 1–9 and also for the so-called leisure time centers attended by approximately 82% of all children aged 6–12. For Years 1–9, there are individual subject syllabuses. In these, objectives and core content, as well as performance standards referred to as ‘knowledge requirements’, are defined, whereas detailed content, materials and methods are to be decided locally.
Personal development dialogues between students, teachers and guardians are to be held at least once a term, and written reports are issued from primary school up to school Year 6 as part of each student’s individual development plan. Formal grades, however, are not awarded until school Year 6, when pupils are around 12 years old. Following the curriculum reform in 2011, there is a rule concerning grades stating that all aspects of the knowledge requirements have to be met for the student to achieve a certain grade; hence, compensation for single weaknesses in individual profiles is not accepted (for example, excellent oral proficiency is not allowed to compensate for a somewhat weaker written proficiency). This is a much criticized phenomenon, which, according to recent political declarations, is likely to be changed. However, the currently existing rules have had a strong impact on teachers’ and schools’ grading practices during the past decade.

1.3.1. English in Swedish Compulsory School

English is one of the core subjects in Swedish schools, studied from early years until the end of upper secondary school (Year 12). It has been a compulsory subject in Swedish comprehensive school since the early 1950s. The starting point was initially school Year 5, but it has since then been gradually lowered (see, for example, Holmstrand 1983). As for today, local school authorities decide about the introduction of EFL instruction, which, however, cannot be later than Year 3 (Sveriges Riksdag 2011). In addition, what needs to be emphasized is that children obviously learn English in school, but the influence of learning outside school, sometimes referred to as extramural learning (Sylvén 2006), is an essential factor to take into account when discussing EFL education and levels of proficiency in Sweden and elsewhere.

Teachers of EFL in Year 6 are often trained as so-called class teachers, that is, teachers who teach a large number of subjects to their students. English in these teachers’ training program was abolished as a mandatory subject in 1987 and was not reintroduced until 2011. Hence, a number of teachers for Year 6 do not have a formal education, including the subject and pedagogy, to teach English. According to statistics from the National Agency for Education (NAE), the percentage of fully qualified teachers of English in school Years 4–6 in 2019 was 62.1%, and 57.3% in 2016 (information for 2013 not available). Among other things, this means that a number of EFL teachers for Grade 6 may not have had a teacher education in which the content and effects of changes in the national syllabus for English were discussed, nor aspects of the language teaching methodology and assessment.

1.3.2. National Assessment of English

There is a long tradition of educational assessment at the national and public level in Sweden. In connection with the current goal- and criterion-referenced system, introduced in the 1990s, national assessment materials of various kinds, formative as well as summative and covering different subjects, are provided throughout the school system. The aims of the tests have varied to some extent over time, but the intention has always been for them to be advisory in the sense that teachers are to combine the results with their own continuous assessments when awarding individual grades. Traditionally, however, albeit much debated and changing gradually, teachers are responsible for marking their own students’ national tests as well as for assigning final grades (for more information, see Gustafsson and Erickson 2018). The responsibility for test development and gradual quality control, as described in a national framework at the system level (Skolverket 2017), lies with different universities, appointed by the NAE.

The national tests and assessment materials for foreign languages are developed at the University of Gothenburg in a standardized collaborative process, including different categories of experts, among which students and teachers have an important function (for further information, including samples of tasks, see the project website; Erickson and Åberg-Bengtsson 2012).

Reflecting the subject syllabuses, part of the national curricula, the typical national test of English comprises three subtests, focusing on: (A) the oral production and interaction;
(B) reception (listening and reading, with a mix of selected and constructed response items); (C) written production and interaction. Tasks are developed in an iterative process, in which small- and large-scale pretesting has a central role. During these iterative rounds, all students and teachers are asked to comment on the different tasks. Teachers are also active in the composition of tests and in standard setting (Erickson et al. 2022). Eventually, following a decision by the NAE, students’ results are aggregated into a single, individual test grade, in which one subtest result below a Pass (but not more than one) can be compensated for by better results on the other two subtests. Results are also presented in broken-down profiles to clarify individual students’ EFL strengths and challenges. The reason for allowing for compensation in the national test is twofold: the dimensionality of language competence and the vulnerability of a single test performance. As pointed out in Section 1.3, however, the current rules for final subject grading do not allow any compensation. This is an issue that is further dealt with in the current article. Extensive guidelines for marking are provided, comprising principles as well as plenty of authentic, benchmarked examples for the three subtests, together with comments related to the national syllabus (the content as well as performance standards). In addition, each test is accompanied by a broad questionnaire for teachers, which forms an empirical basis for the current article.

1.3.3. Students’ National Test Results

Students in Sweden usually perform very well in English. This is demonstrated in national test results for all levels and in subsequent subject grades, awarded by teachers. In addition, at the end of lower secondary school, i.e., school Year 9, there are international survey results demonstrating that Swedish students’ EFL proficiency is at the top in a European perspective (Bonnet 2004; European Commission 2012). No international surveys of EFL proficiency have been created from other levels of the school system; thus, neither for school Year 6 nor for upper secondary levels.

Statistics are available for the national test at the end of school Year 6 (introduced in 2013), based on a mandatory collection of the total cohort, i.e., all students’ results. Examples of information generated are that the English tests consistently generate the highest results, whereas the results for Mathematics are clearly lower than those for English and Swedish L1.

1.3.4. Teachers’ Use of National Test Results for Grading

As previously mentioned, the national tests in Sweden are advisory in the sense that teachers are to combine the results with their own continuous assessments. The extent, or weight, of this advice has been discussed since the introduction of the system in the 1990s, with periods of worries for too much, as well as for too little, influence on teachers’ grading (Gustafsson and Erickson 2018). At present, in 2022, there is a clear movement towards more standardization both regarding test development, as manifested in the system framework for national tests (Skolverket 2017), and the use of results. Regulations regarding the latter are published in a document labelled ‘General advice’, which, in spite of its title, is to be seen as a strong declaration of intent from the central educational authority. Consequently, from a very general wording about support, the instructions are now that the national test results shall be ‘taken into special consideration’ when final grades are awarded (Skolverket 2018). In connection with this, and also with an upcoming introduction of digitalized tests, models for the external marking of national tests are being developed. Reports from the National Agency for Education, published regularly since 2007, show that there is considerable variability in the use of national test results, with similarities as well as differences between subjects (Skolverket 2007, 2020). In school Year 6, the large majority of students in all subjects (between 68% and 76%) are awarded a final grade identical to the aggregated test grade. As shown in Table 1, the proportion between lower and higher final grades (FG) than test grades (TG), however, differs considerably.
Table 1. Test grade (TG) in relation to teachers’ final grades (FG); school Year 6 (2013, 2016 and 2019).

<table>
<thead>
<tr>
<th>Subject</th>
<th>FG &lt; TG</th>
<th>FG = TG</th>
<th>FG &gt; TG</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>N = 90,176/100,423/108,168</td>
<td>27/21/17</td>
<td>69/74/76</td>
<td>4/6/6</td>
</tr>
<tr>
<td>Mathematics</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>N = 90,285/101,029/108,642</td>
<td>20/7/4</td>
<td>69/73/75</td>
<td>11/20/22</td>
</tr>
<tr>
<td>Swedish L1</td>
<td>%</td>
<td>%</td>
<td>%</td>
</tr>
<tr>
<td>N = 83,167/91,031/94,559</td>
<td>11/10/10</td>
<td>73/68/72</td>
<td>17/23/18</td>
</tr>
<tr>
<td>Swedish L2</td>
<td>-/7/9</td>
<td>-/68/73</td>
<td>-/25/18</td>
</tr>
<tr>
<td></td>
<td>N = -/8991/12,239</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As can be noted, teachers of English often award a lower final grade than indicated by the test grade, whereas the opposite is true for the other core subjects. The differences between Mathematics and English are especially conspicuous. In 2019, for example, roughly three fourths of the students in both subjects were awarded the same final grade as test grade, but in English, 17% obtained a lower final grade as compared to 4% in Mathematics, whereas the opposite was true for the higher grades; in English, 6% was awarded a higher final grade than test grade, as compared to 22% in Mathematics. These tendencies are stable across years and form the starting point for the study reported on in the current article. Regarding English, it should also be pointed out that the tendency of awarding a lower final grade than test grade is unique for school Year 6. In Year 9, there is usually an even balance between those who obtain a lower and a higher final grade, and the tendency in upper secondary school is rather a slight imbalance in favor of a higher final grade.

2. Materials and Methods

2.1. Teacher Questionnaires

The main source of data for the study of teachers’ perceptions and use of national test results of English for school Year 6 was the extensive teacher questionnaires accompanying the tests and were answered anonymously. These questionnaires were standardized to a large extent, within and between the different national tests of EFL, and, thus, comparable over time.

A typical teacher questionnaire for EFL tests comprises, approximately, 35 questions, about the test as a whole and each subtest individually. Responses are sought, for example, regarding relevance in relation to the national syllabus, aspects of content and difficulty, rating procedures, degree of agreement between the test result and teachers’ continuous assessment and the perceived usefulness of the test results for grading purposes. A majority of the questions are of a selected response type (either multiple choice or Likert scales); however, they are frequently accompanied by space for open comments. Until 2012, the questionnaires were paper and pencil-based; after that, online forms were gradually introduced, to be fully implemented by 2016. A distinct decrease in reporting rate could be noticed for all subjects since 2016. An example for English, school Year 6, connected to the current study, was that in 2013 a total of 1042 questionnaires was sent in, the corresponding numbers in 2016 and 2019 being 394 and 501, respectively. Whether, for EFL, this can only be associated with the transition to digital reporting has not been systematically investigated. In the instructions for the tests, teachers are asked to use the questionnaires (answered anonymously) to contribute to test development. The collection and analyses of the forms are handled by the universities developing the tests and results are published, both per se and included in the annual reports for each national test (for further information about the EFL test for school Year 6, see the project website).
2.2. Analytical Methods

To answer the research questions in the current study, two types of questions were used, namely, three with fixed responses and three asking for teachers’ open comments. The former concerned:

- Teachers’ general appreciation of the test (‘It is a good test’);
- Degree of support for grading (‘The test provides good support for grading’);
- Correspondence with teachers’ own assessments (‘The results correspond to my own assessment of individual students’ competences’).

In these questions, the responses were asked for on four-point Likert scales, where teachers were asked to indicate the extent of their agreement (very large, fairly large, fairly small, very small). Given the type of data, with varying response rates between years and obvious self-selection of respondents, the results for these questions were accounted for using mainly standard descriptive statistics, such as frequencies and distributions, and also comparisons within and between the different questionnaires.

To approach the research questions targeting teachers’ articulated opinions about the test, the open comments were focused upon. A first selection was conducted based on the questionnaires that included comments on two specific questions, namely, regarding opinions on the test as a whole and on the part targeting written production (Part C). These two questions were chosen based on long experience in the test developing group and also on different interactions with teachers over the years, in which general opinions on perceived test quality together with comments on students’ written performances were those that consistently conveyed the most information. In addition, this was thought to be especially adequate in relation to the current research questions that aimed to better understand teachers’ perceptions and use of the test results.

The number of questionnaires with comments on the two questions was 236 in the 2016 test and 255 for the 2019 test. Since the number of submitted forms, in general, was much larger in 2013, the number of commented forms obviously also exceeded the other two years. Therefore, it was decided to determine a selection by mechanically deleting every third of the commented forms for 2013 to achieve roughly the same number as for 2016 and 2019. This generated a number of 251 questionnaires with open comments on the two questions. When this was completed, the 2013 sample was checked for different, unintended selection effects; however, it was found to be largely comparable to the two other groups and, thus, kept intact. Consequently, as shown in Table 2, the threefold dataset for the analyses of teachers’ comments was the following:

<table>
<thead>
<tr>
<th>Year</th>
<th>Test as a Whole</th>
<th>Part C</th>
<th>Total Number of Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>2103</td>
<td>142</td>
<td>251</td>
</tr>
<tr>
<td>2016</td>
<td>117</td>
<td>119</td>
<td>236</td>
</tr>
<tr>
<td>2019</td>
<td>151</td>
<td>104</td>
<td>255</td>
</tr>
<tr>
<td>Tot n</td>
<td>410</td>
<td>332</td>
<td>742</td>
</tr>
</tbody>
</table>

The analyses of the comments were conducted using systematic thematic analyses, including rounds of iterations. In this, the two researchers worked independently, basically following the different stages defined by Braun and Clarke (2006): familiarization with the data; generating initial codes; searching for themes; reviewing themes; defining and naming themes; producing the report (the latter meaning discussions and joint, tentative decisions determined by the researchers). In addition, external validation was conducted by an experienced language teacher and test developer, who analyzed, in total, 120 comments (16.2% of the material) that were randomly chosen to represent reactions to the two questions (the whole test and Part C) and the three years involved. In this, the thematic categorizations tentatively agreed upon by the researchers were used and commented upon.
3. Results

In the following, based on the three research questions (RQs) defined in Section 1.1, the results of the analyses of teachers’ responses to the fixed-answer questions about the test and their open comments from the teacher questionnaires for 2013, 2016 and 2019 were presented. In addition, the external validation was briefly described. What needs to be borne in mind is that the results were not automatically generalizable, mainly due to the apparent self-selection in the material.

3.1. Research Question 1—General Attitudes

The first research question focused on teachers’ general attitudes to the national test:

What are Year 6 teachers’ responses to fixed-answer questions about the national tests of English?

To answer RQ 1, three questions of a general character in the teacher questionnaires were studied, focusing on the degree of general appreciation of the test, perceived support for grading, and correspondence with own assessment. In this, all submitted questionnaires were included. As for the degree of general appreciation for the three years, the following results emerged (Table 3):

Table 3. Teacher responses to the statement ‘The test as a whole is good’ (2013, 2016 and 2019).

<table>
<thead>
<tr>
<th>Year</th>
<th>Agree Completely</th>
<th>Agree to a Large Extent</th>
<th>Agree to Some Extent</th>
<th>Do Not Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013 (n = 1042)</td>
<td>49%</td>
<td>48%</td>
<td>2%</td>
<td>0%</td>
</tr>
<tr>
<td>2016 (n = 394)</td>
<td>61%</td>
<td>37%</td>
<td>2%</td>
<td>1%</td>
</tr>
<tr>
<td>2019 (n = 501)</td>
<td>68%</td>
<td>29%</td>
<td>3%</td>
<td>0%</td>
</tr>
</tbody>
</table>

As shown, attitudes to the test in general were very positive, with only 2–3% of the respondents agreeing only to some extent or not at all with the Likert statement “The test as a whole is good”. One difference between the years was noticeable, however, namely, regarding the proportion of ‘Agree completely’ and ‘Agree to a large extent’. Here, the responses from the 2013 group were equally divided between the two categories, whereas the 2016 and 2019 groups more frequently chose the most positive alternative. Furthermore, a gradual increase in ‘Agree completely’ responses was noticeable.

Another aspect of teachers’ general attitudes to the national test concerned the role and weight of the test, more precisely, the perceived degree of support for grading. The following reactions were reported (Table 4):

Table 4. Teacher responses to the statement ‘The test as a whole supports final grading’ (2013, 2016 and 2019).

<table>
<thead>
<tr>
<th>Year</th>
<th>Agree Completely</th>
<th>Agree to a Large Extent</th>
<th>Agree to Some Extent</th>
<th>Do Not Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013 (n = 1042)</td>
<td>50%</td>
<td>45%</td>
<td>4%</td>
<td>1%</td>
</tr>
<tr>
<td>2016 (n = 394)</td>
<td>51%</td>
<td>43%</td>
<td>4%</td>
<td>1%</td>
</tr>
<tr>
<td>2019 (n = 501)</td>
<td>67%</td>
<td>25%</td>
<td>7%</td>
<td>0%</td>
</tr>
</tbody>
</table>

The responses to this question were quite similar to the ones regarding general appreciation, in the sense that only a few (5–7%) seemed less satisfied with the support they obtained from the test for their grading decisions. A marginal increase in this category could be seen for 2019; however, it was too small to be interpreted in any substantial way. As for the two positive options, taken together, a large majority of the respondents chose one of these, with an increase in the most positive option (‘agree completely’) for the 2019 test.
The third question aimed to capture teachers' attitudes to the tests, focused on the perceived degree of correspondence between the national test results and their own assessment. As shown in Table 5, the responses generated the following results:

Table 5. Teacher responses to the statement ‘To what extent do you feel that the test results coincide with your own assessment of individual students’ competences?’ (2013, 2016 and 2019).

<table>
<thead>
<tr>
<th>Year</th>
<th>Large Extent</th>
<th>Fairly Large Extent</th>
<th>Fairly Small Extent</th>
<th>Very Small Extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013 (n = 1042)</td>
<td>37%</td>
<td>57%</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>2016 (n = 394)</td>
<td>48%</td>
<td>48%</td>
<td>3%</td>
<td>1%</td>
</tr>
<tr>
<td>2019 (n = 501)</td>
<td>51%</td>
<td>46%</td>
<td>3%</td>
<td>0%</td>
</tr>
</tbody>
</table>

In this case as well, the attitudes expressed indicated satisfaction. The vast majority of the respondents declared that the national test results coincided with their own assessments to a fairly large or large extent, and only 3–6% felt that the results deviated considerably from their own continuous observations. What should be noted, however, is that the balance between the two most positive options was fairly even; hence, it seems quite clear that there were things that differed between the two sources of support for grading.

In conclusion, Year 6 teachers’ responses to the general, fixed-answer questions about the national test were distinctly positive regarding the general appreciation, degree of support for grading and correspondence with their own assessments.

3.2. Research Question 2—Teachers’ Comments

The second research question concerned teachers open comments on the national test as a whole and the part focusing on written competence.

What are Year 6 teachers’ open comments on the national tests of English, focusing on the test as a whole and on written production?

As described in Table 2, a total of 742 teacher comments, emanating from questionnaire responses for three years (2013, 2016 and 2019), was analyzed. Of these, approx. 55% dealt with the test as a whole, and the remaining 45% with Part C of the test, focusing on written competence. In analyzing this, the two researchers worked independently, with one interim discussion of observations and a final meeting, where results were compared and three tentative themes agreed upon.

In the iterative readings, a wide array of comments related to different aspects of the test was identified, some examples being the content of tasks, issues of time, rating guidelines and individual benchmarks, but also construct- and system-related opinions regarding, for example, the national language syllabus underlying the test, the model for aggregating subtest results into a single test grade, often compared to the current regulations for final subject grades (Section 1.3.2). The interim meeting between the researchers revealed considerable coherence and agreement regarding the interpretation and categorization. Based on this, further strengthened by the continued readings, eventually, three overarching, partly overlapping themes were tentatively identified in the comments on the whole test as well as on Part C, namely, comments related to the Content, Rating and Consequences, the first connected to the underlying construct, the second to the actual marking and required levels of competence and the third to different opinions related to effects of the use of the national tests, for both students and teachers. Due to the character and length of the comments, more than one theme sometimes occurred in the same comment. In addition, an evaluative dimension was evident, leading to a classification of each comment as positive, neutral or negative. These three themes and the evaluative dimension were then used by the researchers individually to categorize all comments. The outcome of this showed that comments on content and rating were equally frequent; however, those on content dominated in the responses to the whole test question, the rating-related comments being
most frequent in response to the Part C question. Comments on consequences were, on the whole, less frequent and almost exclusively given in response to the whole test question. As for the evaluative dimension, comments expressing critical opinions dominated, in particular, regarding Part C (the writing test), and were strongly dominated by aspects of rating. The agreement between the two researchers in this part of the analyses was strong, with correlations (Pearson) ranging from 0.82 to 0.96.

To further investigate the tentative themes identified, an external validation was determined by an experienced test developer with a language teaching background in lower secondary school. In total, 120 comments from the whole material (16.2%) were randomly chosen to represent the two questions (whole test and Part C) and the three years involved. After a thorough introduction, including information about the iterative categorization process, the analysis was undertaken using the three tentative themes, Content, Rating and Consequences, as well as the evaluative dimension, Positive, Neutral and Negative. In addition, the external evaluator gave frequent written comments on the different examples. Analyses of the outcome showed that the degree of agreement between the external validator and the researchers was high, with an overall agreement of 87%. Some tendencies could be noted, namely, that the external validator more frequently used a combination of two themes to classify single utterances and, also, that more often the researchers chose the theme Consequences. After some discussions of this, a few minor modifications were determined, and it was then concluded that the themes were a valid means to categorize the comments given by teachers in the three questionnaires.

An observation determined in the analyses of comments concerned certain proportional differences over the years regarding aspects focused upon, or, put differently, that single aspects were clearly more emphasized some years than others. The clearest examples are shown in Table 6 (raw numbers).

Table 6. Variation of aspects emphasized in teachers’ open comments on the whole test and Part C in the 2013, 2016 and 2019 questionnaires (number of comments/raw numbers).

<table>
<thead>
<tr>
<th>Aspect</th>
<th>2013 (n = 251)</th>
<th>2016 (n = 236)</th>
<th>2019 (n = 255)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good test</td>
<td>96</td>
<td>48</td>
<td>41</td>
</tr>
<tr>
<td>Demands too low/leniency</td>
<td>29</td>
<td>60</td>
<td>38</td>
</tr>
<tr>
<td>Aggregation into test grade</td>
<td>22</td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>Time consuming</td>
<td>21</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

As can be noted, positive comments on the tests were frequent in all questionnaires; however, distinctly more so in 2013, the first year of the Year 6 national test. In addition, teachers that year commented more on the time required to administer the test—or rather, all national tests for Year 6—than in later questionnaires. Comments on the perceived leniency, i.e., that demands were too low, generally, but in particular for the Pass level, were considerably more frequent for the 2016 test, whereas a large number of teacher responses in 2019 focused on the aggregated test grade, or rather, the way in which it was constructed, including its consequences for students as well as for teachers (Section 1.3.2).

In addition, a certain degree of invariability in comments was evident between the three years studied. A persistent and fairly frequent comment was, for example, some teachers’ wish to have structured scoring rubrics, often referred to as ‘matrices’, to facilitate the marking of productive and interactive tasks (speaking and writing). Infrequent comments concerned issues of adaptation and accommodation (for example, regarding students with dyslexia), and also explicit requirements for more grammar/focus on accuracy. Some aspects were hardly ever mentioned, for example, students with another L1 than Swedish, and issues related to gender.

Finally, to provide a certain concretization of the data, the following examples of teacher comments (Table 7) illustrate the three categories, including the evaluative dimension.11 What needs to be pointed out is that, given the definition of effects of use for
students and teachers (Section 3.2), there were very few positive comments to be found in the theme Consequence.

Table 7. Examples of teacher comments related to the themes Content, Rating and Consequence, including an evaluative dimension (2013, 2016 and 2019 Year 6 EFL questionnaires).

<table>
<thead>
<tr>
<th>Table</th>
<th>Whole Test (Year)</th>
<th>Part C (Year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>• A nice and varied test that assesses students’ English language proficiency in a good way, based on the national syllabus (2019).</td>
<td>• A good test—many students managed the task without problems—good to start from themselves—good task/theme—nothing complicated but something that everyone can get started with (2019).</td>
</tr>
<tr>
<td></td>
<td>• I think that in general there are too many multiple-choice questions in both reading and listening comprehension. The tasks that students have to answer in their own words and sentences say more about their knowledge (2013).</td>
<td>• The subject made the content superficial and thin, which meant that the students used a simple and everyday vocabulary (2013).</td>
</tr>
<tr>
<td>Rating</td>
<td>• We have coordinated subtests B and C with three other schools, which means that you have not marked your own students’ tests. The coordination also means that the assessment is discussed “across school boundaries”, which we think has been good (2019).</td>
<td>• The richness of variation in the student examples gives good indications. For example, it is not obvious that a correct but meaningless text lacking fluency should be awarded an A. Also, that texts at A level do not need to have completely perfect grammar and spelling. Communication is in focus (2013).</td>
</tr>
<tr>
<td></td>
<td>• When you compare with national tests for other subjects, it still feels as if English is too lenient. Students who have quite considerable problems in everyday teaching still manage a Pass in reading and listening comprehension (2013).</td>
<td>• Very low requirements according to the benchmarks. I’m looking for a matrix! It takes a very long time to get acquainted with all the examples (2016).</td>
</tr>
<tr>
<td>Consequence</td>
<td>• The English test is fun because the results are good and I note that I have been stricter in my assessment of the students than you are and that is good!! (2016).</td>
<td>• Many students have “surpassed” themselves and reached significantly higher results on the national test than in the subject English. Is this a result of my requirements being too high or are the test requirements too low? I think the requirements of the test are too low, because I know how high the requirements are both in years 7–9 and in upper secondary school. Then the national test in year 6 becomes misleading when the requirements for all grades are a little too low (2016).</td>
</tr>
</tbody>
</table>
3.3. Research Question 3—Teachers’ Use of National Test Results

The third research question focuses on teachers’ use of the results from the national test when awarding final grades of English.

What indications regarding the use of the national test results for the final grading of English can be traced in the teachers’ responses and comments?

As shown in the Introduction, there was a clear and systematic discrepancy between the aggregated test grades for English and the final subject grades awarded by teachers, with teachers being considerably harsher when it came to summarizing individual students’ levels of competence than what was indicated by the national test results (Section 1.3.4; Table 1). There was a number of clear indications given, both in response to the fixed-answer questions and in the open comments, which could be used to approach the possible reasons behind the tendency to ‘downgrade’ competences in the final subject grade. One example could be found in the question about the grade levels and benchmarks for Part C (writing), where in one of the questions, teachers were asked about their reactions to the benchmarks illustrating different grade levels. The following table (Table 8) gives an example of this, with E as the Pass level and A as the highest grade level. Aiming to provide further contextualization, the same information was also given for the other productive/interactive subtest in the national test, namely, Part A, focusing on speaking.

Table 8. Teacher responses to the statement ‘The assessment of the examples/benchmarks is . . . ’ (Part C; Part A; 2013: n = 1042; 2016: n = 394; 2019: n = 501).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade E: too low</td>
<td>18%</td>
<td>38%</td>
<td>28%</td>
<td>E: too low</td>
<td>14%</td>
<td>13%</td>
<td>12%</td>
</tr>
<tr>
<td>Grade A: too low</td>
<td>12%</td>
<td>26%</td>
<td>21%</td>
<td>A: too low</td>
<td>11%</td>
<td>7%</td>
<td>8%</td>
</tr>
</tbody>
</table>

A certain degree of fluctuation between the three years is clear for the writing test, with a peak for 2016, whereas the attitudes to the speaking test are more stable. It is also obvious that the overall reactions to the benchmarks for the writing test are more critical than those for the speaking test (with the exception for 2013, which was the year when the national test for school Year 6 was introduced). In addition, the Pass level is usually the most questioned, although a number of teachers also consider the assessment of the highest grade level too low. The following examples of teachers’ comments to the writing test (Table 9) were chosen to illustrate the critical attitudes brought forward:

Table 9. Examples of teacher comments criticizing the assessment of writing/Part C (2013, 2016 and 2019).

<table>
<thead>
<tr>
<th>Teacher Comments</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think that the assessment of the sample texts requires too little of the students and that, in a way, it counteracts my own assessment.</td>
<td>2013</td>
</tr>
<tr>
<td>I think the E-level is set very low. Students in Sweden generally have very good knowledge of English as they come into contact with the English language a lot. This should be observed when assessing.</td>
<td>2016</td>
</tr>
<tr>
<td>I think it is too easy to pass the writing part in the English test. The student performances that accompany the assessment instructions are to some extent too leniently assessed. If you are to get a high grade, I think you should be able to spell, write, explain, etc. without any errors or occasional errors that are not repeated.</td>
<td>2019</td>
</tr>
</tbody>
</table>

As for the open comments to the whole test, they only rarely approached the final grading issue explicitly. However, there were plenty of comments that could be related to
Aspects of use of national test results in a broad sense. In this, content-related comments were often positive, whereas those focusing on rating and consequences were usually much more critical or negative. Additionally, a relatively large number of comments focused on what was perceived as a clash between the national test grade, applying a compensatory model, and the general rules for final grades, where compensation was not allowed (Section 1.3.2). The following examples (Table 10) were chosen to exemplify this.

**Table 10. Examples of teacher comments with a possible relation to final grading (2013, 2016 and 2019).**

<table>
<thead>
<tr>
<th>Teacher Comments</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>It is surprising and gratifying that the students’ performances in English are that good. I probably demand more of them than what is stated in the knowledge requirements, but at the same time I understand that this is the result I have to relate to. It looks as if the tests are too simple, but instead I think they show that students have very good knowledge of English in general in Sweden.</td>
<td>2016</td>
</tr>
<tr>
<td>Think it’s completely crazy with the test grades. It is not reasonable that we think in two different ways. Difficult for some students to understand.</td>
<td>2019</td>
</tr>
<tr>
<td>The overall test is in a way more comprehensive for the subject than the corresponding test [. . . for another subject . . . ], but as the assessment recommendations unfortunately come across as unnecessarily low, the results are somewhat misleading and can therefore also be counterproductive to one’s own assessment. The signals that are sent out rather seem to undermine my “professional” ability to assess.</td>
<td>2013</td>
</tr>
</tbody>
</table>

As shown by the different quotations, opinions regarding the use and implications of the national test results were often contradictory, which was discussed in the following section, as were the other results reported.

**4. Discussion**

The current study set out to investigate an aspect of the national assessment system that could be considered both general and specific, namely, teachers’ use of national test results when awarding final grades. More specifically, the consistent pattern among EFL teachers in school Year 6 to assign final grades that were considerably lower than what was indicated by the national test results were focused upon. The latter phenomenon was observed and reported on by the NAE for a number of years but, to our knowledge, had not been closely studied from the point of view of teachers’ underlying perceptions, interpretations and reasoning. As for the special report on grading practices in school Year 6 from the Schools Inspectorate (Skolinspektionen 2020, Section 1.2), the phenomenon of discrepancy between national test grades and final subject grades was mentioned in general terms, but not commented on from the angle of specificity and systematicity.

The three research questions asked in this study all focused on teachers’ views on the national test and the usability of the results for the purpose of assigning final grades of English in school Year 6. The following discussion was structured based on the three categories of comments emerging from teachers’ open responses, also reflected in the fixed-answer questions (Section 3.2). It was concluded with some overall observations and reflections. What needs to be borne in mind, is that respondents and responses were anonymous; hence, analyses based on personal and/or contextual background variables were not possible.

**4.1. Aspects of Content and Construct**

As mentioned in the Introduction, a communicative and functional approach to language competence has characterized the Swedish national language syllabuses since the early 1980s. Since assessment at the national level aims to reflect the regulatory documents,
the latter, thus, serving as the construct, an action-oriented view of language was also visible in the national tests that comprised tasks focusing on oral and written receptive, productive and interactive competences (Érickson 2020). Generally speaking, teachers seemed to appreciate this, clearly manifested in the responses to the fixed-answer question focusing on a general liking of the test, where almost all respondents (97–98%; see Table 3) agreed completely or to a large extent with the statement “The test as a whole is good”. This was also reflected in teachers’ open comments, where a fairly general liking—“Good test”—was the most frequent comment taken together for all the three years studied; however, this was most prominent for 2013, when the test was first introduced (Table 6). Regarding the latter, it was obvious that many teachers were positively surprised, especially since the test accompanied the introduction of subject grades in school Year 6, previously introduced in Year 8. This was an issue causing much discussion and considerable worry and stress among many teachers, who most often had no experience of, or training in, assigning grades. A number of positive comments also referred directly to the national syllabus and the role of the national tests in their interpretation and operationalization of these documents. In addition, students’ reactions were often mentioned in connection with positive comments on the test. Students being positive to the tests was clearly a benefit from the teachers’ point of view. Negative comments about test content were obviously also given; however, there was no clear pattern, but was rather spread across different aspects, for example, concerning topics, format, time, level of difficulty, etc.

There were, however, also a number of comments demonstrating a considerable questioning of the construct, i.e., the action-oriented view of language articulated in the national syllabus. Quite frequently, this was also clarified in comments about the performance standards, i.e., the requirements for a Pass, as well as for higher grades. A question that may be raised in connection with this is whether attitudes of this kind, to some extent, can be related to the lack of English as an academic subject in teachers’ education, accompanied with no, or very little, training in language pedagogy and grading (Section 1.3.1). According to Shulman (1986), three types of knowledge are essential in teachers’ professional practices, namely, content knowledge, pedagogical content knowledge, including curricular knowledge, and general pedagogical knowledge. Having had a teacher education with little or none of the two first aspects may affect attitudes to language competence as well as to evaluating this competence. It may also be one of the reasons why teachers in school Year 6 handle students’ aggregated test grades differently as compared to EFL teachers for older students, in school Year 9 and upper secondary school (Section 1.3.4). Another reason indicated in the current study was that of consequences, which were dealt with below.

4.2. Aspects of Rating

Comments on levels of difficulty were, to some extent, related to the content and construct, but obviously also had a clear connection to issues concerned with rating. As shown in the initial fixed-answer questions, the vast majority of the respondents (93–95%) declared that the national test gave strong support for their decisions on final subject grades for English (Table 4). In the same vein, between 94% and 97% thought that the test results coincided with their own assessment of individual students’ competences (Table 5). In parallel to this very positive picture, however, a substantial number of critical comments were given on issues related to rating, both the actual marking process and the levels of requirement based on the national performance standards (‘knowledge requirements’). One example of this was the demand for explicit rating rubrics, referred to as ‘matrices’, to handle qualitative analyses of written (and sometimes spoken) texts, or, more precisely, models for aggregating observations of different features into a single evaluative judgement. Here, comparisons were often described with the corresponding national test of Swedish, where such rating tools are offered. The method used in the English tests builds on a combination of the performance standards accompanied by a number of commented benchmarks, consistently using the terminology from the national regulatory documents and focusing on a variety of phenomena common in foreign language written production. Some teachers
were very positive to this, emphasizing aspects of clarification and usability/transferability, while others were not, pointing to the risk of subjectivity and arbitrariness. What was generally clear in the comments, though, was that there was a large number of issues that were perceived as problematic to handle, for example, the comprehensibility and text length in relation to accuracy, different types of texts representing the same overall, qualitative level, and issues of possible weighting of different aspects of language. Some teachers also wanted a larger number of benchmarks, hoping that would provide stronger support for their assessment.

The most frequent comments on aspects of rating, however, focused on the uneven proficiency profiles and the compulsory model for the aggregation of sub-results into a composite test grade (Section 1.3.2). As mentioned previously (Section 1.3), there is a rule concerning final grading saying that proficiency profiles have to be even if a student is to be awarded a certain grade, i.e., all aspects of the performance standards have to be met. This means that you are not allowed to compensate for certain weaknesses with strengths in other areas. The fact that this rule, which in itself is heavily criticized, concerns final grading and not the national tests as such is far from generally accepted, and causes much discussion and criticism that has to do with consequences in a broad sense, thereby closely connected to the crucial aspect of validity (Messick 1989).

4.3. Aspects of Consequences

The criticism of the aggregated test-grade in relation to the final grading was expressed in different ways, most of which indicated frustration with systems that were perceived as contradictory and seemed to clash. Comments on this most often focused on consequences at three levels: an individual teacher level, a pedagogical level, and a structural level.

The individual level concerns what teachers face in situations when they may have to explain to students and caregivers that a student’s aggregated test grade is a Pass, due to the compensation of a single aspect in their uneven proficiency profile, but that the teacher awarded final subject grade will still be an F, based on the teacher’s continuous observations and assessments and following the rules for final, summative grading. In addition, some teachers also feel that the discrepancy between the two systems influences their autonomy and professional authority in a negative way.

The pedagogical level of consequence mainly concerns individual students’ assumed, future difficulties to live up to the subject requirements in the school years ahead, i.e., at higher levels in school, a phenomenon not uncommon in education.

Finally, comments expressing criticism at a structural level dealt with what is often characterized as contradictory and illogical: how can uneven profiles, with strengths and weaknesses, be accepted in one case but not the other? This can also be assumed to be related to teachers’ professional ambition to do their job well, one aspect of which is to follow curricular rules.

What was noticeable in the material analyzed, i.e., in teachers’ comments on the consequences of national test results, was that there was hardly any discussion of what may be seen as a—possibly ‘the’—central issue, namely, how uneven profiles should be regarded and handled in relation to the overarching concept of language competence, or, put differently, in relation to interim, learner language, which is what the subject grade is intended to capture. Instead, different aspects of handling the practical situation and its consequences were often focused upon.

4.4. Overall Observations and Reflections

The material available for the current study offered a wealth of interesting information; however, it was obviously not possible to account for more than just a part, in an article of the current kind. Therefore, a short section comprising overall observations and reflections seemed called for, in order to briefly mention additional interesting aspects and alternative angles to the phenomenon in focus. First of all, however, it needs to be remembered that there were definite limitations to the study, for example, the fact that there were no
background variables available for the respondents. This, in combination with self-selection in answering the questionnaires, emphasizes the need for caution when it comes to drawing definite conclusions or determining strong claims.

An important observation was the apparent contradiction between teachers’ very positive attitudes emerging from the fixed-answer questions and the critical opinions expressed in a number of open comments. Analyses showed that the group of teachers choosing to write personal comments were not quite as positive as the whole group, a tendency that was even clearer for those who expressed very negative opinions, although the differences were small. Consequently, it seemed clear that there was a certain degree of ambiguity among teachers regarding the national test of English—or, maybe, national tests in general.

As shown in the text, some features were variable over time; for others, the opposite was true (Table 6). This may have several reasons, but what seemed quite obvious was that it partly reflected ongoing discussions in society. A clear example of this seemed to be the peak for criticism regarding leniency in the assessment of writing in the 2016 national test, which coincided with intense discussions in teachers’ digital fora, also manifested in the press. Similarly, the focus on rules for and effects of test-grades versus final subject grades in the aftermath of the 2019 test reflected a strong societal debate on the grading system as such (Statens Offentliga Utredningar 2020, p. 43).

Finally, many of the findings in the current study could be related to the discussions on pedagogical and structural issues in the literature within the field of language education. For example, teachers’ struggle with the action-oriented construct has to do with the communicative and action-oriented approach to language and language learning emanating from Hymes (1972) and further developed, for example, by Canale and Swain (1980), and in the work by the Council of Europe on a Common European Framework and its companion volume (2001 and 2020), including extensive materials available on the CoE website, regarding the origin, development and implementation of these documents. In addition, teachers work within a contextual and structural framework, where different factors affect their attitudes as well as their actions, were discussed, for example, by Borg (2003) and Lundgren (1999). The national curricula and syllabuses have an obvious role in this, especially in the use of the regulatory documents, and, hence, in the transition phase between the intended curriculum and the implemented and attained curriculum (van den Akker 2003). Finally, as emphasized, for example, by Kunnan (2004) and Messick (1989), testing and assessments are very much an issue not only of construct coverage, but of ethical use and consequences of results. This is something that has to be learnt, discussed and developed in teacher pre- as well as in-service education, which thereby also connects it to teachers’ professional development within the field of what is often referred to as assessment literacy (Inbar-Lourie 2008; Vogt et al. 2020).

5. Conclusions

The study reported on here was based on local, albeit national, conditions, not automatically transferable to other educational contexts (Dimova et al. 2020). However, in touching upon the multifaceted issue of language education and language assessment and aiming to unfold and better understand teachers’ assessment and grading practices, it was also general and, hopefully, interesting and potentially useful in other contexts. The results were rich, complex and partly ambiguous, clearly indicating a need for continued research in direct collaboration with teachers. In the current case, this was initiated in the form of in-depth interviews with teachers and groups of teachers at different types of schools, aimed to shed further light on the results and issues emerging from the current study.

Afterword

Following a political decision (https://www.riksdagen.se/sv/dokument-lagar/?doktyp=bet&dokstat=beslutad&g=betygs%C3%A4ttning&p=1&st=2) made in February 2022, the rules for grading in Sweden will be changed as from 1 July 2022. This means that
teachers are obliged to make a holistic evaluation of each student’s level of knowledge and award a grade accordingly, i.e., that the rule about all aspects of the formal requirements for each grade level having to be met for a student to receive a certain grade will be abolished. Consequently, compensation between skills will be possible. However, this only applies to grade levels D–A, whereas the original rule remains for the pass level E.

Author Contributions: Writing—original draft, G.E. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The data underlying the study are anonymous and cannot be traced. They consist of responses to large-scale teacher questionnaires from three years, accompanying the compulsory national test of English for school year 6 in Sweden. This test is part of the large-scale, public national testing programme in Sweden, comprising several subjects and taken by whole cohorts of students (>100,000 per year; see further information here https://www.skolverket.se/innehall-a-o/landningsssidor-a-o/nationella-prov, accessed on 20 February 2022).

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 Lower secondary school refers to school Years 4–9 in Swedish compulsory school, and is preceded by a preschool class for 6-year-olds and primary school for school Years 1–3. There is no formal graduation at the end of Year 6, but students often change teachers and sometimes schools.

2 For further information about the Swedish school system at large, see the website of the National Agency for Education (NAE): https://www.skolverket.se/andra-sprak-other-languages/english-engelska (accessed on 20 February 2022).


10 The researchers had partly different roles in relation to the national assessment system, with Erickson representing the perspective of an insider regarding test development and research, and Tholin basically that of an outsider regarding test development, but with long experience as a user of the system and as a researcher. Both authors also have long experience of language teaching and teacher pre- and in-service education, as well as work with language policy issues.

11 Translations from Swedish by the researchers.


References


