National assessment of foreign languages in Sweden: A multifaceted and collaborative venture

Gudrun Erickson
University of Gothenburg, Sweden

Linda Borger
University of Gothenburg, Sweden

Eva Olsson
University of Gothenburg, Sweden

Abstract
The article addresses the local system of national assessment of foreign languages in Sweden, a contextually specific, large-scale system with a summative aim, but also a system aimed to support teachers in their continuous assessment and grading of their students’ competences. In the text, the educational context and the multifaceted nature of national assessment are described and discussed. Furthermore, based on a broad view of validity with use and consequences in focus, different, and partly interwoven, aspects of collaboration in test development are exemplified and discussed, including policy, stakeholders, and research. Special attention is given to contributions by stakeholders, in particular students and teachers. Their involvement is regarded as a central component in the test development process, not only because it widens, deepens, and further develops the competences needed, but also because it increases the possibility to affect and enhance the use of the materials for the justice and beneficence of test-takers and society at large—aspects at the heart of validity. It is emphasized that collaboration requires sensitivity and sensibility from those involved to optimize overall quality and generate reciprocal benefits for all parties.

Keywords
Collaboration, foreign languages, national assessment in Sweden, policy, reciprocity, research, stakeholders, test development, validity

Corresponding author:
Gudrun Erickson, Department of Education and Special Education, University of Gothenburg, Box 300, 405 30 Gothenburg, Sweden.
Email: gudrun.erickson@ped.gu.se
Introduction

Tests obviously differ in various ways, for example, regarding purpose, construct, level of difficulty, intended age group, proficiency level, and format. Regardless of how they may differ, tests share some essential characteristics: They are taken by individuals, who may be considered the most obvious stakeholders, and handled by different actors in terms of interpretation, decisions, and actions. These extensive and varied functions entail a huge responsibility for all parties involved and also emphasize use as a central feature regarding quality. Hence, there is a clear ethical dimension in planning and conducting test development. Inviting stakeholders to contribute, thereby bringing as many perspectives as possible into the process, is one way of approaching the central issue of quality.

This article focusses on national assessment in Sweden. The rationale is to give a close description of a large-scale, public assessment system for foreign languages, embedded in a specific political context, characterized by shared responsibility for quality and fairness of educational assessment between the policy level and the school level. The system includes compulsory national tests whose results are to be combined with teachers’ continuous assessments in their decisions about final subject grades. However, the assessment system also has a formative and pedagogical purpose to support learners in their learning and teachers in their teaching, hence with a clear connection to the classroom level. This local context poses both unique validation challenges and opportunities. With our text, we hope to make a contribution to the test development literature on issues related to what, in our opinion, illustrates a way of resolving some of the challenges of a multifaceted and complex system, namely by involving different stakeholders, in particular students and teachers, focussing on reciprocal benefits for all parties involved.

We base our text on the definition given by Dimova et al. (2020), that local tests are those “whose development is designed to represent the values and priorities within a local instructional program and designed to address problems that emerge out of a need within the context in which the test will be used” (p. 1). In the current case, the participation of different stakeholders, directly affected by the system, is essential to test development and test use. Thus, the conditions favour inclusion, which in itself also reflects the values of the system.

The programme focussed upon is embedded in a specific context, a national educational system with a specific political framework, and therefore not automatically transferable to other contexts. However, as pointed out by Dimova et al. (2020), “even a large, national test could be considered a local test when the values represented by the test reflect distinctive features of a broader instructional context” (p. 2).

After a brief conceptual background and some contextual information, we address the multifaceted nature of cooperation and inclusion within the national assessment system for foreign languages. In this, contributions from students and teachers, most typical of the Swedish assessment system, will be discussed, with interwoven references to policy and research at different stages of test development. Finally, some remarks are given, pointing forward to possible developments of the system.
It needs to be mentioned that the authors of the article to some extent work within the national assessment programme, in slightly different roles. Hence, the perspective given is, to a considerable extent, that of an insider.

**Background**

In the following, a brief background is given regarding conceptual underpinnings, literature within the field, and local contextual circumstances.

**Conceptual considerations and literature**

The development of language assessment materials for the Swedish national system, regarding principles as well as procedures, is conceptually based on a broad view of validity with use and consequences in focus and with a strong emphasis on values (Cronbach, 1971; Messick, 1989a, 1996). This also serves as the point of departure for our text. In addition, Moss’s (1998) focus on conceptual and empirical ways of addressing the consequential aspect of validity in test use has contributed to the approach used in the programme discussed. Furthermore, the broad definition of validity has obvious value-related and ethical implications that can be connected to what Kunnan (2004) refers to as two general principles of justice and beneficence, where justice means that “a test ought to be fair to all test takers; that is, there is a presumption of treating every person with equal respect,” and beneficence that “a test ought to bring about good in society; that is, it should not be harmful or detrimental to society” (p. 33). Other researchers emphasizing the aspect of ethical implications and responsibilities are McNamara (2006) and Shohamy (2001), whose work has been of particular importance to the procedures used in the national assessment programme, not least in further emphasizing a broad definition of validity, including social as well as consequential values. In addition, curriculum theory is of relevance, in particular Van den Akker et al.’s (2003) distinction between three broad categories—the intended, implemented, and attained curriculum. These categories can be connected to different stages of cooperation discussed in the text; with the policy level, responsible for the intended curriculum, and the stakeholder level, in particular teachers for implementation and use of the curriculum, as well as students for the aspect of attainment of curriculum.

Students may be considered the most important group of stakeholders since they take the tests and are distinctly affected by the results. In relation to ethical testing practice, McNamara (2006) pointed to the importance of accountability, that is, “a sense of responsibility to the people most immediately affected by the test, principally the test-takers” (p. 43). Students’ role in assessment is, on one hand, self-evident as they are the ones generating the results. On the other hand, as pointed out by Alderson et al. (1995), students may also contribute more actively in test development and validation processes. Furthermore, issues of test-taker agency, involvement and advocacy have recently been focussed upon, for example, in a symposium during the European Association for Language Testing and Assessment (EALTA) 2021 conference, labelled “Common frameworks and standards in the age of technology-driven language assessment?” (Farrows et al., 2021). In the symposium, the role
of test-takers was discussed, emphasizing the value of “capturing the voices of test-takers” and “making test-takers more central in what we are doing.” As a valid and reliable test should provide an opportunity for all students to demonstrate their proficiency, feedback from test-takers in the development process is essential. In this, with an obvious connection to Kunnan’s (2004) emphasis on fairness, test-taker feedback can provide information about stakeholders’ experiences of completing tasks and tests that would otherwise be difficult to capture (e.g., Erickson, 1999, 2010; Ryan, 2014).

Teachers are another important group of stakeholders and influential participants in test development, as they are the ones directly using the results of tests, not least for assigning grades. Messick (1989b) pointed out that in the validation process, we need to consider “the relevance of the scores to the particular applied purpose and the utility of the scores in the applied setting” (p. 5). Careful analyses of information from teachers about their judgement of the functionality and usefulness of the tasks for the purpose of assessing students’ language proficiency thus make a significant contribution the validation of the materials (cf. Winke, 2011).

The background of national tests

National assessment in Sweden has a long tradition. Final exams and external examiners, were abolished in 1968 and replaced by standardized tests for certain core subjects, for example, English (Marklund, 1987). Initially, these tests were developed by the national educational authorities, but in the early 1980s, the operative responsibility was delegated to different universities with proven competence and experience within different subject domains and in educational assessment, for example, the University of Gothenburg regarding foreign languages (Marklund, 1987). The aim of this transition of responsibility was twofold: first, to increase quality by benefitting from specialized, academic competence, and second, to enhance the perceived legitimacy of the system, the latter due to an ambition to have an external evaluation of the attainment of the goals in the curricula and subject syllabuses. Consequently, using Van den Akker et al.’s (2003) terminology, the intention was to have an independent evaluation made of the intended curriculum in relation to the implemented and attained curriculum.

Two long traditions can be identified: on one hand, a strong trust in autonomous teachers assessing their own students’ competences and awarding final grades used for high-stakes purposes; on the other hand, continuous provision of high-stakes tests at the national level (Skolverket, 2019).2 Reflecting this, since the 1968 reform, all national tests have been advisory, that is, not defined as exams overruling teachers’ decisions but as materials to be combined with teachers’ continuous assessments (Marklund, 1987). In addition, national tests have always been connected to the national subject syllabuses, with a more or less explicit ambition to serve as a tool for clarification and operationalization of the curriculum (cf. Messick, 1996; Van den Akker et al., 2003).

Policy considerations

The national, non-profit assessment system in Sweden is based on political decisions and connected to the national curricula and subject syllabuses, as well as to the prevailing
grading system. The task of handling the system is given by the Ministry of Education to the National Agency for Education (NAE) that, in turn, delegates the responsibility for test development to different universities in the country. The relationship between the NAE and the universities is regulated in annual agreements specifying the assignment, time plan, and budget. The delegation arrangement means a considerable degree of autonomy for the universities involved but also presupposes communication and collaboration. Furthermore, acceptance, in a wide sense, of the national curricula and syllabuses, the latter serving as the construct for the national assessment programme, is self-evident (cf. Van den Akker et al., 2003).

In Sweden, universities are, by definition, state authorities, as is the NAE. The relationship between the two in the national assessment system is not a traditional employer–employee situation but, formally speaking, rather a peer agreement, albeit with differing roles, conceptually as well as operationally. Using Van den Akker et al.’s (2003) terminology, the NAE has the role of implementing the politically agreed upon intended curriculum. The test-developing university level is also part of this implementation and, in particular, the analyses of the attained curriculum, that is, the results demonstrated by students. This multi-faceted relationship requires continuous discussions aimed at finding an optimal balance between administrative and/or political versus academic/research-based concerns. In addition, what may be mentioned is that a system framework for all national tests was published by the NAE in 2017 (Skolverket, 2017) and that principles and procedures from the development of the foreign language tests had a considerable impact on this document, not least regarding the aspect of collaboration.

It could also be mentioned that the relationship between the policy and test development levels is highlighted in the Guidelines for Good Practice developed by the EALTA, where it is stated that “test developers are encouraged to engage in dialogue with decision makers in their institutions and ministries to ensure that decision makers are aware of both good and bad practice, in order to enhance the quality of assessment systems and practices” (section C, p. 3). What could be added is that developing tests to be taken by whole cohorts of students in a country also requires solid knowledge among test developers of the system level, not least regarding issues of curriculum and regulations. Consequently, there is a clear element of reciprocity in the relationship between the policy-making and test-developing agents. It needs to be emphasized that this is something that requires constant attention and a good deal of sense and sensitivity from both sides to optimize the reciprocal value of a broad dialogue, thereby, hopefully, enhancing the quality of the tests.

**National tests of languages**

The national tests build on the national syllabuses for foreign languages, which consequently serve both as the regulatory documents for learning and teaching, and the construct for assessment. As for test content, there is a long tradition of communicatively oriented language assessment, based on the functional description of language competence given in the national curricula since the early 1980s (Canale & Swain, 1980; Malmberg, 2001). Furthermore, the current Swedish national syllabuses for foreign
languages are influenced by, and to some extent also comparable to, the Common European Framework of Reference/CEFR (Council of Europe, 2001), although the seven levels of proficiency defined have not been empirically aligned to the six levels of the CEFR (Erickson, 2019; Erickson & Pakula, 2017). Areas focussed upon are receptive, productive, and interactive competences, as well as intercultural communicative competence (further information in Appendix 1).

Although the assessment materials of foreign languages are of different kinds, they all build on a set of basic principles presented on the project website. These principles emanate from conceptual as well as practical considerations regarding aims, construct, methods, agency, and use (Erickson, 2010, 2018a; Messick, 1989a), summarizing features of processes and products. It is also emphasized that the materials should give test-takers the best possible chances and conditions to show what they actually know and can do with the language in focus (cf. Fox, 2004; Kunnan, 2004).

A typical national test is comprised of three parts: an oral test, in which pairs of students talk individually and discuss different subjects (see Borger, 2018), a receptive sub-test including a listening and a reading comprehension section, with a variety of text types, tasks, and formats, and a writing test in the form of an essay (see Olsson, 2018). There are extensive teacher guidelines for all materials including test specifications, keys to the different tasks, commented responses and authentic samples of benchmarked oral and written performance, cut-off scores, and so on. For reasons of transparency, a substantial number of sample tasks for each test are provided on the web (further information and examples on the project website.)

Cooperation with the main stakeholders

Developing a national test takes approximately two years, during which involvement and cooperation with multiple stakeholder groups take place at different levels, all of whom have expertise in one or more areas of relevance: curriculum developers, teachers with different competences and experiences including special education, teacher educators, and users of the target languages as L1 or L2, as well as researchers within different disciplines (Erickson, 2020). Contact with national and international institutions plays an important role. Among all stakeholders involved, students and teachers have a special role due to their obvious proximity to the materials. Consequently, much attention is paid to bringing these two categories into several stages of the test development and validation phases.

A large part of the test development work is empirical. After construct-based discussions in broad expert groups including teachers, different tasks, items as well as passages and topics for production and interaction, are constructed and successively piloted in small-scale rounds, where the test developers are sometimes present to observe and discuss the tasks with students and teachers. This iterative phase leads to various revisions, and eventually to pre-testing in randomly selected schools in the country, normally comprising a minimum of 400 students per task. In these rounds, anchor items are used to enable comparisons across groups and time. Furthermore, all participating students and teachers are asked to comment on the different tasks. Teachers are also active partners in decisions regarding test composition, as well as in
benchmarking and standard setting (for further information, see the project website, and Erickson, 2018a).

**Contributions of students**

As mentioned, students may be considered the most important group of stakeholders as they take the test and are directly affected by the results. Hence, they provide essential contributions to the test development process (cf. Alderson et al., 1995). One way of involving students in the development of tests is to ask them about their opinions on tasks. Starting in the mid-1990s, systematic incorporation of test-taker feedback became a standard component of the development of foreign language national tests and assessment materials in Sweden, both directly in connection with the piloting of materials, and in feedback forms (see Appendix 2) as part of the large-scale pretesting rounds (Erickson, 1999).

Alongside regular use when analysing and evaluating results and composing tests, student feedback data have also been the subject of several studies, with partly different aims, namely to find out what the students actually notice and comment on, and to see to what extent their retrospective, task-related feedback correlates with their results on the different tasks. Examples of the content related type of studies can be found in Erickson (1999, 2010). Some salient results are that (1) students state they feel positively about the tests, younger students in particular about tasks requiring active language use, spoken as well as written; (2) girls tend to underestimate their achievement when judging retrospectively how well they did on a specific task. This is relatively consistent across proficiency levels and varies only to some extent for different task types. These findings are of immediate value for interpreting and acting on test-taker feedback in test development, for example, regarding content, as well as timing and sequencing of tasks.

Student feedback may also shed light on students’ experiences of different task formats, which is essential information in test development. In Olsson et al. (2018), students’ opinions, expressed on a Likert-type scale in the feedback forms used in large-scale pretesting across 10 years, about three types of reading comprehension tasks were analysed, and the covariance between student views and their scores on the same tasks calculated. The three types of reading tasks were stories with selected and constructed response formats, gap texts with selected response, and gap texts with constructed response; task types regularly included in national language tests. The results showed that students’ opinions about the overall quality of the task, the difficulty of it, and their outcome expectancy after completing it, covaried in a statistically significant way with their scores on the tasks \((r \text{ between } .34 \text{ and } .47^{**})\). The results also revealed that the gap text format with selected response was considered less challenging than the two other task types requiring constructive response, findings of relevance, for example, in sequencing tasks in a test. In addition, following the analysis, Olsson et al. (2018) suggested that further development of the feedback questions could generate more detailed information, for instance related to reasons for considering a task easy or difficult, or for liking or disliking it. Partly emanating from the outcome of the study, the feedback forms
have been revised for the purpose of further capturing students’ opinions and, thereby, enhancing the usefulness of test-taker feedback.

The open comments in the student feedback forms often provide additional information for test development. For example, students’ comments on reading tasks may show when the ambition to base reading tasks on texts that students find interesting has been successful: “This is for real!? I didn’t believe it until I saw the pictures. I love it.” Not all comments are as explicit—sometimes students comment on the content in a way that reveals their engagement in the text implicitly: “I wonder where X is going” or “he must have looked funny.” A comment such as “It was fun to read but it was a little difficult to understand sometimes” indicates that even if the student found the text somewhat difficult, (s)he still thought it was interesting and most likely tried to complete the task.

Sometimes students also provide detailed information on why they felt that they succeeded well or not. The following opposing comments by two students relate to one and the same reading task: “Hard to understand and many difficult words. But I did my best” and “It was very easy to understand and also very easy questions.” Since all students take the same test at a certain educational stage, it is unavoidable that certain tasks are found very difficult or very easy by some students. The last part of the first comment “But I did my best” can be seen as a good sign when test developers analyse the comments; at least this student did not give up even if (s)he was struggling.

As mentioned, it is of interest to compare students’ scores on tasks to their own experience of how well they succeeded. Here as well, the open comments may provide useful information. A student who scored high on a reading task wrote: “The most difficult thing we’ve done so far, at least for me.” Obviously, this student experienced the task as more difficult than what the score revealed. When choosing and sequencing tasks to a full test, the level of difficulty ideally progresses throughout the test, and also within each task. In this, perceived level of difficulty is an important factor to consider and thus student feedback is required.

Furthermore, student comments sometimes demonstrate remarkable insights into what may be perceived as more theoretical aspects, in particular regarding validity, thereby contributing to the conceptualization of test development and interpretation. The following two examples, taken from Erickson (2010), provide examples of two 15-year-old students’ personal phrasing of what Messick (1989a, referring to Cook & Campbell, 1979) identified as major threats to validity, namely construct underrepresentation: “A bad test/assessment is the one which is only about grammar, because if a person knows grammar well, it doesn’t mean he/she can speak the language as well and communication is the most important thing in language study,” and construct-irrelevant test variance: “A bad test is when we just have a little time to do it on. Because then you have to stress thru the test and maybe you cannot show how much you can.”

To summarize, test-taker feedback provides indispensable complementary information about the functionality and acceptance of tasks among students at different levels of proficiency (cf., e.g., Kunnan, 2004; Ryan, 2014). Importantly, it needs to be remembered that this is an area where students are the true experts and where test developers need their assistance to optimize the quality of the materials. Examples of aspects where students’ input plays an essential role in test development and validation concern
in-depth understanding of how various materials function; detection and elimination of possible ambiguity or obscurity in instructions, tasks and items; the choice of topics, texts and tasks, not least with regard to bias; and optimization of sequencing (Erickson, 2010).

**Contributions of teachers**

As the overall purpose of the national language tests is to support teachers’ assessment and grading of students’ language competences, teachers’ acceptance of the tasks and tests as fair, reliable, and valid, in relation to the performance standards stated in the syllabus, is central (cf. Messick, 1989b; Winke, 2011). Collaboration with teachers is done in several ways, both by inviting them to take part in different working groups within the project and through collecting questionnaire responses in connection with pre-testing and regular tests. In these rounds, teachers often provide detailed comments, for instance, about students’ reactions, the relevance of topics, the perceived relation between the tests and the syllabus, the clarity of instructions, and so on. They also offer suggestions on how to enhance the quality and functionality of the tasks.

Results from teacher questionnaires regarding regular tests are made publicly available on the project website, which makes it possible to study possible trends as well as changes of attitudes regarding specific issues. Generally speaking, questionnaire responses reveal that teachers are most often very positive to the national tests of foreign languages regarding content as well as format, and consider them valuable and supportive tools in their grading. Furthermore, they often point out that the tests clarify and operationalize the national syllabuses in a useful way (Skolverket, 2019). Thus, it may be claimed that the national tests have an aligning function between curriculum and classroom (cf. Biggs, 1996), and that the intended curriculum is clarified in relation to the implemented curriculum (Van den Akker et al., 2003). Criticism obviously also occurs, however often focussing on system related issues, for example, workload and current regulations regarding national test results in relation to final grading (Borger, 2019b; Erickson & Tholin, forthcoming; see below).

The teacher questionnaires provide data for in-depth analyses of teachers’ opinions on the national tests and how they are used. For instance, a study by Erickson and Tholin (forthcoming) concerns teachers’ attitudes and actions, based on their comments on, and use of, the national English as a Foreign Language (EFL) test for school year 6 (students aged 11–12). More precisely, the focal point of the study is the fact that, in spite of generally very positive attitudes to the test, teachers at this level award final grades of English that are consistently and significantly lower than indicated by the aggregated test results. As mentioned earlier, the national tests are advisory, which means that no total correspondence is expected between the two, but the strong tendency to “downgrade” their students’ competences, is very different from the three EFL tests targeting higher educational stages, that is, the end of compulsory school, and two levels in upper-secondary school. In the study, a total of 742 teacher questionnaires accompanying the test for 3 years were analysed, both regarding closed and open responses. Results indicate that most criticism concerns two issues, related to the construct and to the system, respectively: requirements are
considered too low for the pass level, in particular concerning Writing, and the principle behind, and effects of, NAE regulations regarding an aggregated test grade are heavily criticized.

As a complement to the teacher questionnaires distributed in connection with the annual administration of all national tests, a study was conducted regarding upper-secondary school teachers’ perceptions and use of the speaking test (Borger, 2019b). An oral test, including peer-to-peer interaction, has been a mandatory part of the national tests of English for more than 20 years. The responsibility for organizing this is delegated by the NAE to the school level. Because teachers are involved as both test administrators and raters, their input on different issues related to implementation contributes valuable information on consequential aspects of test use. In the study, 267 teachers responded to a nation-wide online survey about their administration and scoring practices. Results showed clear variation in how the speaking test was implemented at the school level. This is a direct consequence of the delegation of responsibility to the local level, however, with obvious implications for standardization. Teachers also raised concerns regarding limited resources, most commonly related to administrative support in the organization of the speaking tests and time for collaborative rating with colleagues, the latter recommended by the NAE as a way of further enhancing rater reliability. In addition, responses revealed that teachers were positive to the national assessment materials and found the commented benchmarks and the rater guidelines to be good support for their assessment of students’ oral proficiency. Since the guidelines accompanying the tests serve an important function in standardizing the ratings done by a large and diverse group of teachers, this kind of teacher feedback is essential. As seen in this study and in Erickson and Tholin (forthcoming), teachers’ input on different aspects of test use also highlights system-related issues. Although these cannot be directly addressed in test development, the research findings contribute to ongoing policy discussions at the system level about the format, function and functionality of the national tests.

As teachers in Sweden have a high degree of autonomy in assessing their students’ competences, including the responsibility for rating national tests, issues of rater consistency are of particular interest, especially concerning the performance-based parts of the national tests. Rater agreement among teachers is followed as part of the successive validation of the test materials, both in connection with benchmarking and in post-test studies. For example, in Borger (2018), aspects of rater agreement and decision-making were studied in connection with the speaking sub-test. The study involved 17 national and 14 international raters, a design chosen also to enable an external, tentative, and small-scale rating of Swedish students’ performances in relation to the CEFR for Languages. Raters provided holistic ratings of a sample of six audio-recorded paired speaking performances, taken from a pre-testing round aimed for a national test at upper-secondary level. To explore decision-making, the raters were also asked to write comments on salient features that influenced their scoring. Descriptive analyses of scores illustrated some variability and differences in rater severity, which is in line with previous findings in research on rater effects in second language performance assessment (e.g., Eckes, 2005). Furthermore, although inter-rater consistency between the
Swedish teachers was overall satisfactory (median correlation $r_s = .77; \tau_b = .66$), there was still some room for improvement.

Additionally, content analysis of the raters’ written justifications showed that raters attended to a wide range of students’ oral competences when making their decisions, not least their interactional skills. Although mainly focussing on the same general features of oral proficiency and commenting on them in a similar way, some differences in rater orientations were noted between the Swedish teacher group and the CEFR raters. For example, in terms of interactional competence, the Swedish teachers commented more frequently on the way students made use of topic development moves and interactive listening strategies, whereas the CEFR raters made comparatively more comments on how turn-taking was managed in the conversation (Borger, 2019a). The rater comments also illustrated the complexity involved in rating peer-to-peer interactions, where an individual test-taker’s performance is likely to be influenced by the way the conversation is co-constructed with the other student (cf. May, 2011; Weir, 2005). The results indicate that further guidance and support are needed to apply the standards in a more consistent way, and to handle co-constructed interaction, including interlocutor effects. The latter also implies the need for clarification and elaboration of the national subject syllabuses. Furthermore, although collaborative rating is recommended by the NAE, a more formalized system of double marking would most likely strengthen consistency and agreement. In addition, continuous professional development is essential to help teachers develop a common understanding of the standards.

Teachers also play an important role in benchmarking of the productive and interactive parts of national assessment materials (Erickson, 2010). When, for example, a writing task has been successfully tried out on a large scale, teachers are invited to take part in the rating and benchmarking process. A reference group of experienced teachers and test developers independently rate a large number of student texts before, during a joint meeting, selecting a number of texts as examples of specific grade levels. The selected texts, accompanied by comments relating them to the standards/syllabus, function as benchmarks in the guidelines for the national assessment materials.

Benchmark texts may also be particularly valuable for further analyses as they have been rated, discussed and selected by groups of teachers as typical for a certain proficiency level. In post-test analyses, it is of relevance to explore and compare linguistic features of texts rated at different competence levels. For instance, the EFL syllabus stipulates that the depth and breadth of students’ linguistic repertoires need to be considered in the assessment; hence, vocabulary is an important aspect of the test construct. In a study by Olsson (2018), student texts ($n = 71$) that had been used as benchmarks in national tests of English, ranging from primary to upper-secondary school, were analysed and, using corpus-based methods, compared with special attention paid to vocabulary. The comparisons showed that the lexical scores, for instance related to the frequency level of the vocabulary included in the texts and the variation of vocabulary, overlapped to a considerable extent across educational stages. Texts corresponding to the performance standards for the highest grade level in year 9 could, for instance, score higher with regard to lexical measures than texts corresponding to a pass in the highest course investigated in upper-secondary school. The results thus confirm highly diverse proficiency levels among students, not only between
educational stages but also at the same educational level, posing a challenge for teachers to provide adequate education for all students in a class and for national tests to provide an opportunity for all students to show what they actually can do with their language. As the assessment system, including the national tests, is based on standards for certain educational stages, proficiency levels beyond or below those requirements are not explicitly assessed in the tests, pointing to a possible limitation of the system. However, as mentioned, the purpose of the current national system is to support teachers’ assessment of their students’ language competences at certain educational stages in relation to the standards for the specific school years, all this in a clearly inclusive and comprehensive school system. For other purposes, for instance to describe individual students’ competence levels regardless of educational stage, a more flexible system would be needed.

Furthermore, as shown in Olsson (2018), in-depth analyses of benchmark texts, generated through collaboration with groups of teachers, may provide information of clear relevance for test development, an example being an increasing use of corpus-based methods for analyses of lexical features in the development of new tasks as well as in analyses of student writing (cf. Olsson & Sylvén, 2017).

Collaboration with teachers in test development obviously involves a strong element of reciprocity. Taking part in task development or benchmarking and standard setting processes contributes to teachers’ professional development, including assessment literacy. The response to invitations is almost always positive and teachers often comment on the value of participating in test development. Eventually, it needs to be pointed out that the results of studies such as the ones mentioned here, also reach teachers through teacher and in-service education as well as through publications.

Discussion

Handling the development of materials for a national assessment system entails considerable responsibility and also requires a group of contributors with a wide range of competences. In the programme focussed upon here, it also means developing very different materials, ranging from formative to summative; materials to be used in classroom work, as well as compulsory, high-stakes tests, albeit advisory, taken by almost all students in the country, from young learners to adults.6 As shown, this obviously entails both challenges and opportunities regarding test development and validation. In our text, building on a definition of validity including ethical dimensions and consequences (Kunnan, 2004; Messick, 1989a), we have argued that collaboration is a central component, a backbone, not only because it widens and deepens the competences needed, but also because it increases the possibility to affect and enhance the use of the materials, an aspect at the heart of validity. Moreover, an essential aspect of collaboration is reciprocity, that is mutual benefit for the parties involved, which further emphasizes the ethical dimension (cf. Kunnan, 2004).

Involvement of multiple stakeholders is key to test development. In this text, the focus has been on the main stakeholders, namely students and teachers, as we consider the consistent, multi-faceted and large-scale collaboration with these two categories of experts the most typical and unique for our local context. This is further underpinned by the fact that the degree of teachers’ autonomy, agency and, indeed, responsibility in the
Swedish system is large and that the national tests are aimed to advise them in this and to enhance quality and equity in the system. Here, there is a strong relation to an expanded view of validity, with use and consequences in focus (Messick, 1989a; Moss, 1998), as well as ethical concerns (Kunnan, 2004) and curricular considerations (Van den Akker et al., 2003).

As mentioned, the groups of stakeholders most directly affected by the national tests are students and teachers. Whereas the latter group has a long tradition of active participation in national test development in Sweden, large-scale student participation was introduced in the 1990s (Erickson, 1999) and has undergone gradual development and expansion since then. Students’ input is self-evident from the point of view of validity, comprising aspects of ethics and democracy (cf. Shohamy, 2001), but it is also unique in the sense that only the students themselves can communicate their comprehension, reactions and feelings in a broad sense. This information is essential, especially in an inclusive school system like the Swedish, where the vast majority of students go to school together until the age of 15–16 and take the same national tests. As shown, students’ input provides valuable information about the functionality of the materials, for example, regarding clarity of instructions, bias and choice of topics, and also adds a dimension to the definition of difficulty, namely students’ retrospective, task-related self-assessment (Erickson, 2010; Olsson et al., 2018). In addition, and as an example of reciprocity, a large number of students through the years have expressed that they appreciate being asked to contribute their thoughts and feelings about different tasks and tests, especially since they do not seem to be used to this.

Regarding teachers’ participation in test development, it should be mentioned that the large majority of the test-developers in the national testing project are originally certified teachers of languages with substantial experience from different levels and types of schools. Many of them have added further to their original academic qualifications through courses on language and language education, methodology, and so on, in some cases leading to a PhD. There is also an explicit aim within the project group to communicate rationales for and results from continuous studies as well as regular tests. This is done in reports made public on the websites of the project (e.g., Erickson, 2018b) and the NAE as well as in national and international journals. Furthermore, there is a strong ambition among test developers and researchers in the project to take part in teacher pre- and in-service education, thereby contributing to knowledge and professional development within the field of educational assessment.

As described and exemplified, external teachers take part in different groups within the project. They also play an important role in administering locally the large pre-testing rounds undertaken for all materials, reporting on their various observations and analyses. Thus, taken together, teachers’ role in the development of national tests is multifaceted, with the bridging function between theory and local practice being one of prime importance, and clearly an aspect of validity (cf. Dimova et al., 2020; Moss, 1998; Van den Akker, 2003; Winke, 2011). This becomes very clear in connection with the development of extensive teacher guidelines. These guidelines are seen not only as a self-evident component of a test, but also as a means of further standardizing the rating done by a very large and diverse group of teachers in the country. In addition, they may also have an in-service training function in enhancing language teacher assessment.
literacy, or perhaps rather professional development, in a wide sense (Inbar-Lourie, 2008; Vogt et al., 2020). It may even be claimed that these guidelines implicitly serve the function of rater training in a country where thousands of teachers, more or less experienced, have the responsibility for marking their own students’ national tests and awarding final subject grades.

In studies related to national tests and test development, students and teachers make essential contributions, as indicated by the examples given in this article. The results of such studies have an immediate impact on test development in many cases or, at times, they may identify or point to dilemmas that need to be handled at the policy level.

**What’s in it for whom?**

Finally, as mentioned already in the introduction, reciprocity is an essential aspect of collaboration, that is, mutual benefits for the parties involved. Hence, putting it straight and simple, the question “What’s in it for whom?” is relevant in relation to the involvement of stakeholders described. We hope that this has been clarified in the text, but a condensed summary may be useful:

- **Policy-makers** receive gradual input on the intended curriculum and on various development and information activities from stakeholders as well as from research(ers).
- **Test-developers** receive input, and benefit, from discussions at the policy level as well as from opinions communicated by students and teachers and other stakeholders (e.g., at the school and municipal levels). Students’ input, in particular, can be considered indispensable, due to its unique and irreplaceable character. In addition, research, conducted internally or by external colleagues, obviously contributes to the development process.
- **Students** contribute their reactions and opinions, in itself an awareness-raising and empowering activity. Their opinions are thoroughly analysed and included in test development, thereby adding to the quality of the actual tests taken by, and affecting, all students in the country.
- **Teachers** take an active part in most stages of test development; by administrating, piloting and pre-testing, communicating their experiences, for example, by answering extensive questionnaires; also through work in reference groups, where policy- and construct-related issues are frequently discussed; in task development, when composing tests, and by taking part in rating and benchmarking activities. Furthermore, over the years, many teachers have reported that the tests as such and the accompanying guidelines serve as clarification of the national syllabuses and, to some extent, as in-service education in a system where teachers have a huge responsibility for assessment and grading.
- **Researchers** get invaluable information/data from studies of a wide range of phenomena related to language competences as well as to development and performance processes, and their results feed back into test development, to students and teachers, as well as to the policy level.
It also needs to be emphasized, that involvement and cooperation with the different agents are partly interwoven in dynamic ways, resulting in a truly multifaceted system of collaboration to enhance the quality of the national assessment system. In this, there is also an ethical responsibility for all parties involved of analysing and discussing what is expected and what could and/or should be done. Obviously, this is a complex and demanding task. What should be borne in mind, however, is the ultimate shared goal, namely—in accordance with Kunnan’s (2004) principles of justice and beneficence—an assessment system that is valid and stable, that can be trusted and that, as far as possible, gives students a chance to show what and how much they know and can do with their language(s), at the same time providing teachers with tools that support and advise them in their teaching, assessment and grading.

The Swedish national assessment system is extensive but distinctly local and specific, not least in the dual role of serving both summative and formative purposes within its educational context. Also, it is currently in a phase of distinct change in several ways. Most noticeable is a transition to digital provision of all materials, gradually introduced and intended to be fully implemented in 2026. In addition, the structure and role of the tests, as well as issues of marking and agency, are under investigation. A government decision (July, 2021), stipulating that central rating of all national tests is to be introduced, implicates a substantial change of the current system, as teachers will no longer rate their own students’ national tests. However, details about the practical implementation of the decision have not been presented for the time being. In this, issues of validity and reliability, as well as feasibility, are intensively discussed to find solutions that balance in the best possible ways technical ambitions and solutions and considerations of quality and fairness. In light of this, and in our opinion also transferable to other contexts, the importance of the collaborative philosophy and procedures focussed upon in this article is even further emphasized. The dialogue between the policy and academic levels needs to be intense and mutually respectful, and collaboration with students and teachers and utilization of their experiences and suggestions have to remain strong and be even further taken into account. Moreover, research must be conducted in parallel with planning as well as implementation to provide evidence for use in decisions at different levels (cf. Moss, 1998). Furthermore, basic issues of a more generic kind, not least teachers’ and students’ roles and responsibilities, need to be approached from a multifaceted perspective and in respectful collaboration between different educational levels and actors.

Acknowledgements
The authors would like to thank all students and teachers, who regularly and positively contribute to the development of the national assessment system in Sweden—this article is to a considerable extent based on their contributions. Many thanks also go to other collaborative partners, in and outside the University of Gothenburg, not least our colleagues in the National assessment project (NAFS) for stimulating discussions and constant support. Finally, we gratefully acknowledge the constructive comments and suggestions given by the anonymous reviewers and the guest editors.

Declaration of conflicting interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Gudrun Erickson https://orcid.org/0000-0002-1145-6164

Notes
1. “Foreign languages” refer to modern languages learnt within the school system of a country where the language is not the majority language.
2. Skolverket is the National Agency for Education, the central administrative authority for the public school system (see further https://www.skolverket.se/andra-sprak-other-languages/english-engelska)
3. http://www.ealta.eu.org/guidelines.htm
4. https://www.gu.se/nationella-prov-frammande-sprak/in-english-auf-deutsch-en-francais-en-espanol
5. Examples from students in Grade 9.
6. National tests are not compulsory in special schools for students with learning disabilities; also, exceptions for individual students, based on special circumstances, can be made by school principals.

References
Alderson, C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. Cambridge University Press.
Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher Education, 32, 347–364. https://doi.org/10.1007/BF00138871
Borger, L. (2018). Investigating and validating spoken interactional competence: Rater perspectives on a Swedish National Test of English [Unpublished doctoral dissertation]. Gothenburg University Publications Electronic Archive. http://hdl.handle.net/2077/57946
Borger, L. (2019a). Assessing interactional skills in a paired speaking test: Raters’ interpretation of the construct. Apples—Journal of Applied Language Studies, 13(1), 151–174. https://doi.org/10.17011/apples/urn.201903011694
Borger, L. (2019b). Evaluating a high-stakes EFL speaking test: Teachers’ practices and views. Moderna Språk, 113(1), 25–57. http://ojs.ud.gu.se/ojs/index.php/modernasprak/article/view/4318
Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1(1), 1–47. https://doi.org/10.1093/applin/1.1.1
Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Rand McNally.
Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.
Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd edition, pp. 443–507). American Council on Education.
Dimova, S., Yan, X., & Ginther, A. (2020). Local language testing: Design, implementation, and development. Routledge. https://doi.org/10.4324/9780429492242
Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. Language Assessment Quarterly, 2(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2
Erickson, G. (1999). Från Sp 8 till Äp 9—om utvecklingen av ett nytt nationellt prov i engelska i grundskolan [From Sp 8 to Äp 9—on the development of a new national test of English for lower secondary school]. In Papers on LANGUAGE—Learning, Teaching, Assessment. Festskrift till Torsten Lindblad. Rapport nr 1999:02 (pp. 202–230). Göteborgs universitet, Institutionen för pedagogik och didaktik

Erickson, G. (2010). Good practice in language testing and assessment—A matter of responsibility and respect. In T. Kao & Y. Lin. (Eds.), A new look at teaching and testing: English as subject and vehicle (pp. 237–258). Bookman Books.

Erickson, G. (2018a). Bedömning av språklig kompetens [Assessment of language competence]. In G. Erickson (Ed.), Att bedöma språklig kompetens. Rapporter från projektet Nationella prov i främmande språk [Assessing language competence. Reports from the National assessment project for foreign languages] (pp. 15–40). University of Gothenburg, Reports from the Department of Education and Special Education. http://hdl.handle.net/2077/57683

Erickson, G. (2010). Good practice in language testing and assessment—A matter of responsibility and respect. In T. Kao & Y. Lin. (Eds.), A new look at teaching and testing: English as subject and vehicle (pp. 237–258). Bookman Books.

Erickson, G. (2018a). Bedömning av språklig kompetens [Assessment of language competence]. In G. Erickson (Ed.), Att bedöma språklig kompetens. Rapporter från projektet Nationella prov i främmande språk [Assessing language competence. Reports from the National assessment project for foreign languages] (pp. 15–40). University of Gothenburg, Reports from the Department of Education and Special Education. http://hdl.handle.net/2077/57683

Erickson, G. (2019). Holistic peer analyses of National Tests in relation to the CEFR. In A. Huhta, G. Erickson & N. Figueras (Eds.), Developments in language education: A memorial volume in honour of Sauli Takala (pp. 49–66). http://www.ealta.eu.org/resources.htm

Erickson, G. (2020). National assessment of foreign languages in Sweden. https://www.gu.se/nationella-prov-frammande-sprak/in-english-auf-deutsch-en-francais-en-espanol/information-in-english

Erickson, G., & Pakula, H.-M. (2017). Den gemensamma europeiska referensramen för språk: Lärande, undervisning, bedömning—ett nordiskt perspektiv [The common European framework of references for languages: Learning, teaching, assessment—A Nordic perspective]. Acta Didactica Norge, 11(3), 1–23. https://doi.org/10.5617/adno.6434

Erickson, G., & Tholin, J. (forthcoming). Overall a good test, but. . . —Swedish lower-secondary teachers’ perceptions and use of national test results of English.

Erickson, G. (Ed.). (2018b). Att bedöma språklig kompetens. Rapporter från projektet Nationella prov i främmande språk [Assessing language competence. Reports from the National assessment project for foreign languages]. University of Gothenburg, Reports from the Department of Education and Special Education. http://hdl.handle.net/2077/57683

Farrows, M., Harding, L., Whitelock, D., & Harsch, C. (2021). Common frameworks and standards in the age of technology-driven language assessment? Keynote symposium at the 17th EALTA conference. http://www.ealta.eu.org/documents/conferences/2021/EALTA%202021_Programme_updated.pdf

Fox, J. (2004). Biasing for the best in language testing and learning: An interview with Merrill Swain. Language Assessment Quarterly, 1(4), 235–251. https://doi.org/10.1207/s15434311laq0104_3

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. Language Testing, 25(3), 328–402. https://doi.org/10.1177/0265532208090158

Kunnan, A. (2004). Test fairness. In M. Milanovic & C. Weir. (Eds.), European language testing in a global context. Proceedings of the ALTE Barcelona conference, July 2001 (pp. 93–115). Cambridge University Press.

Malmberg, P. (2001). Språksynen i dagens kursplaner [Languages in today’s course plans]. In R. Ferm & P. Malmberg (Eds.), Språkboken: En antologi om språkundervisning och språklin- lärning [The language book: An anthology on language teaching and language learning] (pp. 16–25). Skolverket/ Liber Distribution.

Marklund, S. (1987). Skolsverige. Del 5: Läroplaner [School Sweden, Part 5; Curricula]. Liber.
Appendix 1

Table 1 shows the estimated relationship between the seven levels of foreign language proficiency defined in the Swedish national curricula, including national tests and assessment materials, and the CEFR, based on textual analyses (cf. Erickson, 1999, 2019; Erickson & Pakula, 2017). “Year” refers to school year in lower secondary school, “Course” to the upper-secondary school. Nb. The Swedish pass levels represent minimal requirements. Also, it needs to be emphasized that English has a clearly different role than other modern languages, not least due to the degree of exposure in society. In addition, English is compulsory throughout the school system and introduced earlier (in school year 3, at the latest). Other modern languages—traditionally French, German and Spanish—are optional in lower secondary school, usually starting in school year 6, and mandatory only in academically oriented study programmes in upper-secondary school (for further information, see Tholin, 2019).

Table 1. National foreign language levels and tests/assessment materials in Sweden.

| Level Swedish curriculum Pass level | Estimated CEFR level | English | Modern Languages |
|------------------------------------|----------------------|--------|-----------------|
| 1                                  | A1.2                  | Diagnostic materials Years 1–6 | Year 9; “Students’ Choice” (third foreign language; starts in Year 8; taken by < 2%) Course 1 |
| 2                                  | A2.1                  | Year 6 National test | Year 9; “Language Choice” (starts in Year 6; started by c. 85 %, completed by c. 70 %) Course 2 National assessment materials |
| 3                                  | A2.2                  | Diagnostic materials Years 7–9 | Course 3 National assessment materials |
| 4                                  | B1.1                  | Year 9 National test | Course 4 National assessment materials |
| 5                                  | B1.2                  | Course 5 National test | Course 5 |
| 6                                  | B.2.1                 | Course 6 National test | Course 6 |
| 7                                  | B.2.2                 | Course 7 National assessment materials | Course 7 |

CEFR: Common European Framework of Reference.
Appendix 2

Example of TTF (test-taker feedback) form for reading comprehension

We need YOUR help to make really good tests of English!
Please react to the following statements about Back to Nature (BtN) and then write your comments

|   | Yes, absolutely | No, absolutely not |
|---|----------------|-------------------|
| 1. BtN was a good test |   |   |
| 2. It was difficult |   |   |
| 3. The text was interesting to read |   |   |
| 4. I learnt something from doing this type of test |   |   |
| 5. There were many words that I didn’t understand |   |   |
| 6. BtN tested something that is important |   |   |
| 7. I have done tests like BtN before |   |   |
| 8. I think I did well on BtN |   |   |

Comments about Back to Nature (you can write in English or in Swedish)

---

Many thanks!