Evaluating a high-stakes EFL speaking test: Teachers' practices and views

LINDA BORGER
University of Gothenburg

Abstract
In the present paper, teachers' implementation practices and views of practicality regarding a paired speaking test, part of a high-stakes national test of English as a foreign language (EFL) in the Swedish upper secondary school, were investigated. In Sweden, national tests are centrally developed but internally marked by teachers at the schools where they are administered. Two-hundred and sixty-seven teachers participated in a nation-wide online survey and answered closed and open-ended questions. The responses reflect how teachers implement and perceive the national speaking test in relation to purposes and guidelines. Furthermore, challenges relating to the implementation were also reported. The results showed that there were variations in how the national speaking test was implemented at the local school level. This has clear implications for standardisation, but must be considered in relation to the decentralised school system that the test is embedded in, which requires local decisions to be made and local responsibility to be taken. In addition, many teachers perceived that they did not receive enough support from the school management, indicating that clearer routines and administrative support are needed. Statistical tests were undertaken to explore potential differences related to certain background variables. It was found that school size accounted for some of the variation in teachers' responses, with teachers at smaller schools perceiving the practical implementation of the oral tests to be more problematic and time-consuming. The paper concludes with a discussion of the implications of the findings for the practice of high-stakes speaking assessment programs, focusing on the educational context of the current investigation.
Keywords: Swedish national test of English as a foreign language (EFL), paired speaking test, practicality, test administration, validation

1 Introduction

Sweden is one of few countries in Europe where teachers are entrusted with marking students' achievement on high-stakes national tests independently without external supervision (European Commission/EACEA/Eurydice, 2009). In a highly decentralised educational system, where great trust is placed in teacher assessments, teachers are undoubtedly an important stakeholder group that can provide valuable information concerning *contextual* (Weir, 2005), as well as *consequential aspects* of test interpretation and use (Messick, 1989; Moss, 1998; Ryan, 2002). Despite this, teachers' views are rarely included in validations of language tests (Norris, 2008).

In the present paper, the practical implementation of a paired speaking test, part of a high-stakes national test of English as a foreign language (EFL) in the Swedish upper secondary school, is explored from the teachers' point of view, who are also the administrators and examiners of this test. As pointed out by Fulcher (2003), the

practice of test administration and the impact of the environment in the context of speaking assessments is an underexplored area, leading him to conclude that "the local conditions in which speaking tests are held are worthy of serious consideration" (p. 155).

When looking at the European school context, a European Commission report on national testing of foreign languages in Europe (European Commission/EACEA/Eurydice, 2015) concluded that speaking was the least tested skill in the 34 European countries included in the study, whereas reading was most commonly assessed: "It is probable that the complexity of testing speaking skills as well as the high costs involved, mean that this skill is either simply not tested, or that the speaking tests are designed at school level instead of centrally" ("Highlights Report: Languages in Secondary Education," 2015, p. 2). In light of this, the national EFL speaking tests in the Swedish context are particularly interesting to investigate, as they are standardised tests (see, e.g., J. D. Brown, 2018), which are centrally developed but administered and internally marked by teachers at the local school level.

It is worth pointing out that the Swedish educational system with teacher-based assessments of national tests is highly debated, both nationally and internationally (Nusche, Halász, Looney, Santiago, & Shewbridge, 2011; OECD, 2015). Re-markings of samples of national tests have been undertaken by the Swedish Schools Inspectorate, pointing to lack of inter-rater reliability for the performance-based parts (see, e.g., Swedish Schools Inspectorate, 2017), although it should be acknowledged that the methodology used by the Schools Inspectorate has been questioned (Gustafsson & Erickson, 2013). However, the oral parts of the national tests have not been included in the re-markings as recordings are not mandatory, which adds to their interest.

2 Background

2.1 Practicality as an aspect of validity

Although "not a quality of the assessment itself, but rather of the entire process of assessment development and use" (Bachman & Palmer, 2010, p. 261), practicality is viewed as an integral part of validity in prominent language test validation frameworks (Jin, 2018). For example, Bachman and Palmer (1996) proposed the notion of *test usefulness* as an overarching concept in place of construct validity, including six 'test qualities': *reliability*, *construct validity*, *authenticity*, *interactiveness*, *impact*, as well as *practicality*, which fills the function of prioritizing the investigations of the qualities. Whereas the first four qualities address test score interpretation, the latter two refer to consequential aspects of test use. Practicality concerns the implementation of language tests and is defined as the relationship between "the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities" (Bachman & Palmer, 1996, p. 36). In other words, "practicality is a matter of the extent to which the demands of the particular test specifications can be met within the limits of existing resources" (Bachman & Palmer, 1996, p. 36).

Resources are classified into three types: (a) *human resources*, for example test administrators and administrative support, (b) *material resources*, including space, equipment and materials, and (c) *time*, for example with respect to administering the test.

Furthermore, Weir's (2005) socio-cognitive framework of language test validation emphasises that both the internal, mental processing of the test taker (the *cognitive* dimension), and the context in which the test task is performed (the *social* dimension) need to be taken into consideration in the evaluation of language tests. *Context validity* refers both to the representativeness of test tasks in relation to the target language use situation, as well as "the conditions under which the task is performed arising from both the task itself and its administrative setting" (Weir, 2005, p. 19). Although Weir (2005) emphasises that practicality is "not a necessary condition for validity" (p. 49), practical aspects such as the setting and test administration are viewed as "primary considerations affecting validity" (p. 82). Another central component of the framework, with implications for practicality, is *scoring validity*, which encompasses all aspects of the rating process, including rating conditions.

2.2 Teachers' involvement in high-stakes speaking assessments

Xerri and Vella Briffa (2018) point out that teachers' involvement in high-stakes language tests, "including policy-making, design, development, implementation, rating, moderation, and training" (p. 2) is an underresearched area, despite the fact that high-stakes language tests have grown in importance internationally. Previous research in both general education and language assessment has highlighted unintended negative outcomes of high-stakes testing, such as negative effects on teaching as well as on student and teacher motivation (Cheng, Watanabe, & Curtis, 2004; Jones, 2007; Winke, 2011). Evidence has also been presented concerning unreliability and bias of teachers' assessment (Harlen, 2005). On the other hand, it has also been argued that teachers' involvement in high-stakes testing can lead to positive outcomes (Black, Harrison, Hodgen, Marshall, & Serret, 2011; Harlen, 2005; Popham, 2009). For example, within the context of language testing, positive effects include teacher empowerment, engagement in professional development and increased assessment literacy (Xerri & Vella Briffa, 2018).

As pointed out above, studies on teachers' involvement in high-stakes speaking tests are limited; however, a few examples can be mentioned. Winke (2011) surveyed 267 teachers and school administrators about their views of the administration of an English language proficiency test in Michigan, USA. The findings, based on an exploratory factor analysis, and thematic analysis of open-ended comments, indicated that respondents were generally pleased with the impact of the exam; however, teachers were apprehensive about the effectiveness of the administration, especially in terms of the logistics of administering the speaking component of the exam and the large amounts of educator time required for this.

Furthermore, East (2015) conducted a national survey to investigate teachers' (N = 152) views on the relative usefulness (Bachman & Palmer, 1996) of a newly

implemented high-stakes assessment of foreign language spoken proficiency in schools in New Zealand. Teachers were asked to compare the new speaking test – *interact* – consisting of a portfolio of peer-to-peer interactions, with the earlier model – *converse* – which took the form of a one-time summative teacher-led interview. In general, teachers found *interact* to be a more valid and authentic representation of students' spoken proficiency. However, concerns were raised regarding the practicality and fairness of collecting ongoing peer-to-peer performances. East concluded that there was a tension between the test developers' ambition to use a dynamic assessment format and fundamental notions of standardisation and reliability.

A survey conducted by Sundqvist, Wikström, Sandlund, and Nyroos (2017) targets the national EFL speaking tests in the Swedish educational system, focussing on the teacher as examiner of standardised tests. Sundqvist et al. (2017) collected data from 204 school teachers at the lower secondary school level and conducted 11 interviews with the aim of examining teachers' practices and views regarding aspects of test administration. The findings indicate that teacher practices differed greatly, which, according to the authors, has negative implications for standardisation and raises doubts about the summative function of the test. It was therefore recommended that the test authorities frame the test as non-standardised and emphasise its formative qualities.

Additionally, in the context of the French Baccalauréat, Bellhouse (2018) investigated a convenience sample of eight English secondary school teacher examiners' views of the addition of a new foreign language speaking component, consisting of both a monologue and an interview-type interaction. The teachers strongly believed in the value of the new speaking component and reported that students had increased their attention to the speaking construct as a result. However, they were clearly concerned with the lack of training and resources provided by the school/Ministry. Bellhouse (2018) concluded that "assessment literacy training should be included in the professional development of teachers, especially when they assume the role of examiners for national High-Stakes language tests" (p. 85).

Finally, it is worth mentioning that annual questionnaires are conducted with teachers who administer and mark the national EFL tests in the Swedish school context[1]. Results indicate that during the past ten years, more than 95% of teachers have expressed positive views towards the national EFL tests, both to the principle of national testing as such and to the support for grading provided in the assessment materials (Erickson, 2017). With regard to the speaking component, teachers are positive towards the paired test format in terms of students' opportunities to display their speaking ability and its close alignment with the action-oriented view of communication (Council of Europe, 2001) expressed in the foreign language syllabuses. The criticism given mainly concerns work load (Erickson & Åberg-Bengtsson, 2012).

[1] The results are published on the National Assessment Project webpage: https://nafs.gu.se/prov_engelska/engelska_gymn/resultat.

In sum, the reviewed studies show that teacher perspectives of high-stakes speaking assessments contribute important information regarding validity aspects, such as construct representativeness, impact and consequences. Additionally, there are strong indications that practicality and standardisation are two issues that require particular attention with regard to high-stakes speaking tests.

2.3 The paired speaking test format

There are three predominant speaking test formats traditionally employed in educational contexts: (1) *monologue*, for example in the form of speeches, oral presentations and story-telling, (2) *interaction with an examiner*, often in the form of a structured one-to-one interview, and (3) *interaction with one or more test-takers*, including interactive tasks such as role plays and conversations (see O'Sullivan, 2013, for an overview of speaking test methods). In the Swedish national EFL speaking test, a peer-peer interaction format is used.[2]

There are many advantages of using this format in a school context. To start with, it is more time efficient to conduct speaking tests in pairs or groups rather than individually. There is also the potential of a positive washback effect (Messick, 1996) as the test format may encourage interactional speaking tasks in the language classroom. In addition, it has been demonstrated that paired and group speaking tasks offer opportunities for candidates to display a wide range of language functions, particularly interactional skills (Brooks, 2009; ffrench, 1999; Kormos, 1999; Lazaraton, 2002), which are not as easily elicitable from the examiner-led interview format. However, there are also challenges. A major concern relates to *interlocutor effects* (O'Sullivan, 2002), in other words how an individual test-taker's performance is "affected by the way the discourse is co-constructed by the person they are interacting with" (Weir, 2005, p. 153). Various interlocutor characteristics and their effect on discourse and scores have been investigated, for example *proficiency level* (Csépes, 2009; Davis, 2009; Iwashita, 2001; Nakatsuhara, 2006; Norton, 2005), *gender* (O'Loughlin, 2002; O'Sullivan, 2000), *personality* (Berry, 1993, 2007; Nakatsuhara, 2009; Ockey, 2009), and *acquaintanceship among interlocutors* (O'Sullivan, 2002). However, the findings are inconclusive and appear to be highly context-dependent. Nevertheless, it is clear that the matching of candidates needs to be carefully considered in this test format.

2.4 Aim and research questions

In light of the scarcity of research on teachers' involvement as administrators and assessors in the context of high-stakes speaking tests, the main aim of the present study was to provide a stakeholder perspective of the national EFL speaking tests by exploring self-report data from upper secondary teachers of English in Sweden regarding their *implementation practices* and *views of practicality*. The following research questions are addressed:

[2] The speaking component with a peer-peer interaction format became a mandatory part of the national test battery in 1998 for compulsory school and in 2000 for upper secondary school.

- How do teachers implement the national EFL speaking tests in the Swedish upper secondary school?
- What are teachers' views regarding the practicality of the national EFL speaking tests and what potential challenges do they identify?
- Do teacher background variables, more specifically gender, teaching experience and the size of the school, relate to their practices and views of the national EFL speaking tests?

3 Context of the study

The current study focuses on the national EFL speaking test at the upper secondary level in the Swedish educational system. The primary function of the Swedish national assessment system is to enhance comparability and equity within the school system. The national tests are not final exams but have an advisory function in teachers' decision-making regarding students' final grades and should be used in combination with teachers' continuous assessment.[3] Since the national test results are consequential for students' final grades, which are used for selection to higher education, the tests are regarded as distinctly high-stakes.

The Swedish national tests of foreign languages, just like the national tests in other subjects, are centrally designed and developed following rigorous rules for standardisation[4] (J. D. Brown, 2018; Erickson & Åberg-Bengtsson, 2012). However, they are marked by teachers, who are provided with detailed test instructions and guidelines, as well as commented samples of benchmarked performances, in addition to the national standards. Typically, there are three subtests: a speaking test, a writing test, and a section focusing on reception, i.e. listening and reading comprehension. Since Sweden has a highly decentralised school system (Ahlin & Mörk, 2008), the responsibility for the implementation of the national speaking tests is entrusted to the head teacher who should plan the organisation together with his/her staff at the local school level. In order to create good conditions for a fair and reliable scoring, peer marking, or co-rating, i.e. a process whereby two or more teachers collaborate in the rating procedure, is recommended but not regulated.

The purpose of the speaking test is to test oral *production* and *interaction* (Council of Europe, 2001); in other words students' ability to communicate effectively in spoken English. The test task consists of a conversation in which students should speak about, develop their thoughts on, and discuss a given topic,

[3] It is worth noting that as from 2018, the weight of the aggregated national test results in relation to teachers' grading has been strengthened. According to the revised Education Act (SFS 2017:1104), teachers shall 'pay special attention' to the national test results.

[4] For more information on the development process of the Swedish national tests of foreign languages, see Erickson and Åberg-Bengtsson (2012).

on their own and in interaction with others. [5] Test instructions stipulate that two students, or possibly three, should take the test together. The students have 15 minutes preparation time before the test, and the total time allowed for the speaking test is about 15 minutes.

4 Data and methods

4.1 Data collection procedure

The study reported here is part of a larger survey investigating teachers' views of *test usefulness* (Bachman & Palmer, 1996) in relation to the national EFL speaking test in the Swedish upper secondary school. In the present paper, the focus is on one specific validity aspect, *practicality*. Data for this study were collected through an on-line survey administered to teachers of English at upper secondary schools in Sweden during spring 2017. The sampling frame was created on the basis of a database, compiled by *Statistics Sweden*, with information on all Swedish upper secondary school units, at the time of the current study 912 in number (excluding adult education). Given the target population size, a sample size of 150 schools was deemed appropriate. Simple random sampling was used to select 150 schools from the sampling frame. Selection parameters included school size (>100 students), school type (approximately 70% public and 30% independent schools, which is close to the national distribution), program (>50 students at programs preparatory for higher education, where national tests are compulsory to a larger extent) and regional spread.

The invitation for the survey was sent via email to the administration and head teacher of the 150 upper secondary schools with a request to forward it to all English teachers at their school. The survey was open for two months. Two reminders were issued, which resulted in 267 individual responses, thus meeting the desired response rate of >200.

4.2 Participants

Of the respondents, approximately 75% were female. The average age was 47, ranging from 26 to 68 years. The participating teachers had taught for an average of 16 years (range 1–42, *SD* = 10). As regards teacher certification, a majority of the respondents reported being certified EFL teachers (96%).

The survey was anonymous, but as part of the demographic information collected, respondents were asked to provide the name of their school, which 95% did. It could be concluded that responses had been obtained from at least 119 of the 150 schools in the sample; corresponding to a response rate of 79% at the school level. The number of individual responses from each school varied from one to eight teachers. Of the 119 schools, 77% were public schools, i.e. run by the

[5] On the National Assessment web page, sample tests are provided for reference: https://nafs.gu.se/prov_engelska/exempel_provuppgifter.

municipality, and 23% independent schools, i.e. organised and owned by a company, a foundation or an association. In this respect, the obtained sample seems representative of the national composition, where, in 2017–2018, approximately 26% of all students at the upper secondary school attended an independent school and 74% a public school (Holmström, 2018). Furthermore, the 119 schools were representative in terms of geographic spread; Sweden is administratively divided into 21 counties and all were represented. With regard to the group of 31 non-response schools, the composition was similar to that of the 119 schools from which responses had been obtained, both in terms of distribution between independent and public schools and geographic spread, indicating no obvious non-response bias.

4.3 Survey instrument and data

The questionnaire was constructed by the researcher and built on two sources: (1) test specifications and guidelines for the national EFL speaking tests (Swedish National Agency for Education, 2016a), and (2) the framework of test usefulness outlined in Bachman and Palmer (1996). The questionnaire was pre-tested and modified in two steps; first, feedback was given by five topic experts, before it was piloted with a group of in-service teachers. The final survey included 60 items and was divided into four sections addressing (1) implementation practices, (2) assessment in relation to national regulatory documents and the purposes of the test, (3) perceptions of test content and format, and (4) demographic characteristics of respondents. Item formats included both closed-ended questions (Likert-scale items and selected multiple-choice items) as well as open-ended questions that gave respondents the opportunity to comment on a selection of the closed questions. In this paper, a subset of items focussing specifically on teachers' *implementation practices* and their *views of practicality* were examined (See full list of examined items in Appendix A). The questions in the survey were optional and the response rate was generally very high (> 95%).

For the purposes of this study, three background variables were examined in order to find out whether teachers' practices and views of practicality differed with respect to (a) gender, (b) years of teaching experience, and (b) the size of the school where the respondent worked (two variables). Gender was chosen as a background variable since research on oral language testing has suggested that characteristics of the examiner, such as their first language (L1), gender, personality and communication style, may influence discourse and assessment results (Amjadian & Ebadi, 2011; A. Brown, 2003; Reemann, Alas, & Liiv, 2013; Winke, Gass, & Myford, 2013). In addition, teaching experience is one of the factors commonly taken into consideration when investigating teacher quality (e.g., Rice, 2010; Wiswall, 2013) and teachers' practices and attitudes towards standardised testing (e.g., Urdan & Paris, 1994), although it has been suggested that the largest gains occur in the first five years of teaching (Harris & Sass, 2011). Teaching experience in the Swedish educational system is also closely related to rating experience, as teachers are usually involved in marking national tests every year. The last background variable was chosen since it was hypothesised that the local conditions

and environment for teachers as assessors of large-scale speaking tests are slightly different depending on school size.

Gender was based on teachers' self-report of their sex (male/female). Years of teaching experience was based on teachers' self-report of how many years they had been working as a teacher (< 5 years / 6–10 years / 11–20 years / > 20 years). Two background variables were used to indicate the size of the school: (1) the number of English teachers at the school where the respondent worked, based on self-report (1–5 teachers / 6–10 teachers / 11–40 teachers) and (2) the number of students at the school where the respondent worked, also based on self-report (<500 students / 500–1000 students / 1001–1500 students / > 1500 students).

4.4 Data analysis

The data for the current study consisted of both quantitative data from the responses to closed-ended questions in the survey and qualitative data from the open-ended questions. To give a deeper understanding of teachers' implementation practices, the quantitative analyses are illustrated through examples of teachers' open-ended comments. The statistical analyses are based on descriptive statistics and tests of associations with background variables. The alpha level was set at $p = 0.05$. Pearson's chi-square test was used to test the association between background variables and categorical items. The strength of association between *gender*, consisting of two categorical groups, and ordinal dependent variables was measured using the Mann-Whitney U-test.[6] To examine the association between continuous independent variables (teaching experience in years and size of school measured by the number of English teachers and the number of students) and ordinal dependent variables, Spearman's rho correlations were used. SPSS Statistics, Version 25.0 (IBM Corp., 2017) was used to compute the statistical analyses.

5 Results

In the following section, the results of the survey will be presented in relation to four aspects of the practical implementation of the national EFL speaking test: (a) *administration,* (b) *scoring,* (c) *availability of resources* and (d) *perceived practicality*. Finally, the association between background variables and teachers' practices and views will be explored.

5.1 Administration

Teachers' responses regarding the administration of the speaking tests revealed some variation in practices (see Table 1, Appendix B). To start with, test instructions state that it is optional to administer the speaking tests successively during the designated test period or during a shorter period of time. Almost half of the respondents (49%) answered that they administered the speaking tests during *a shorter period of a couple of days* (Q2*a*), whereas 40% administered the test *successively during the test period*. Regarding timing (Q2*b*), a majority of the

[6] The Mann-Whitney U-test is a nonparametric alternative to the independent samples t-test.

teachers administered the tests *during lessons* (61%), as opposed to *outside lessons* (12%). Nearly one third of the teachers (27%) combined the two. In the open-ended comments, many teachers remarked that it was very time-consuming to administer the speaking tests during their regular English lessons, and they were concerned that this took time from teaching. Teachers working at schools where the tests were centrally organised and scheduled seemed more positive about this solution, as illustrated in the teacher comment below:

> When the oral part of national tests was carried out during lesson time, it took up unreasonably much time. At our school, we have therefore switched to concentrating the speaking tests to two days. This arrangement is better overall. [Resp. 124][7]

As regards recording (Q2*c*), which is recommended in the test guidelines, the results showed that nearly half of the teachers in the sample (49%) *recorded* the oral tests, whereas about 40% reported that they *did not record* the tests. The main reason mentioned for not recording was lack of time for re-listening. It was also argued that recording was not necessary when two teachers conducted the speaking tests together. Some teachers thought recording might feel stressful for students, thus inhibiting their performances. However, the opposite was also pointed out:

> Recording is necessary for a reliable assessment, I believe. That is why I always use it, unless a student opposes, which has never happened so far, on the contrary, most students seem to feel it is reassuring. [Resp. 70]

In terms of grouping (Q2*d*), the results revealed that the majority of the teachers in the sample administered the test *in pairs* (42%), whereas a third of respondents (29%) used the response option *in groups*, and another third (30%) divided students into *both pairs and groups*. 154 respondents provided an open-ended comment to clarify how many students they generally included per group. A majority (72%) stated that they administered the test with groups of a maximum of three students, whereas the rest reported using groups with up to four students. As can be seen, the recommendation to use a maximum of three students was not followed in all cases, which may be explained by the fact that the test instructions were phrased somewhat more liberal with respect to the number of students in each group previously.

In the instructions for the national speaking tests in the upper secondary school, no specific guidelines are given on the matching of students.[8] However, in more general terms it is emphasised that an important aspect of the test is that the individual student feels that he/she is given the opportunity to display his/her full ability, which a careful matching of students might contribute to. Open-ended

[7] All examples from the open-ended comments have been translated from Swedish.

[8] It should be noted, however, that such advice exists in the test guidelines for the lower secondary levels, where it is stated that, in most cases, it is not suitable to match students who are at very different proficiency levels, or students who, for various reasons, do not get along well together, since this might affect assessment results negatively.

comments on grouping practices were made by 30% of the teachers. The most commonly mentioned considerations when matching students were that they should have similar proficiency levels and communication styles, and that they should feel comfortable with each other, which could reduce test-taker anxiety. The open-ended comments clearly showed that the matching of students was an essential task to teachers.

Another issue concerns teacher intervention. Teachers are instructed to keep in the background of the conversation and let the students control the conversation. However, the teacher should also "encourage students to give each other roughly equal speaking opportunity" (Swedish National Agency for Education, 2016a, p. 17). Results from questions 6*a* and 6*b* revealed some variation in teachers' answers. Almost half of the teachers replied that they *often* or *always* (47%) intervened if one of the students did not get enough speaking opportunity. A less pronounced inclination to intervene (28%) could be noticed if the students 'got stuck' in the conversation. In the open comments, many teachers pointed out that students generally managed to solve problems in interaction on their own, without teacher intervention. This was attributed to the fact that students were used to practising communicative strategies in the classroom, implying a positive washback effect.

5.2 Scoring

The next section deals with teachers' reported scoring practices. The respondents were asked to indicate the degree of support for rating they had from the various assessment materials available, as shown in Table 1.

Table 1. Frequency of item 17* (*N* = 267)

Assessment materials	Valid *N*	Minimum	Maximum	Mean	SD
a) analytic assessment factors	264	1.0	7.0	5.29	1.4
b) benchmarked and commented samples of performances	265	1.0	7.0	5.28	1.7
c) national performance standards	264	1.0	7.0	5.11	1.6
d) commentary materials for English	261	1.0	7.0	5.03	1.6

**"To what degree do you believe you have support from the following assessment materials when scoring the national speaking tests?", 7-point scale from 1=Minimal support to 7=Very large support*

In general, teachers found the assessment materials to be of good support. The analytic assessment factors (*M* = 5.29, *SD* = 1.4) and the benchmarked and commented samples of oral performances (*M* = 5.28, *SD* = 1.7) were perceived most favourably, whereas the national performance standards for oral production and interaction and the additional English subject commentary materials provided

by the Swedish National Agency for Education (2011) had slightly lower means. However, the standard deviations indicate variation in teachers' opinions.

Whereas the assessment factors (Appendix C) are analytic and focus on different qualitative aspects of spoken production and interaction, the national performance standards are expressed in the form of a holistic rating scale with performance descriptors for the different grade levels (Appendix D). The teachers thus seemed to favour the analytic factors in terms of support. In this regard, it is interesting to note that a report by the National Agency for Education (2016b) shows that Swedish teachers find the holistic national performance standards to be unclear, and their usefulness has consequently been questioned.

As regards the benchmarked samples of oral performances, teachers were also asked to indicate how adequate they thought the grading was on a three-point categorical scale (1=too lenient, 2=adequate, 3=too severe) (Q20). Overall, the participating teachers reported that the grading of benchmarked performances at the higher grade levels, C and A, was adequate (89% and 86% respectively). However, for the lowest passing grade, E, one fourth of teachers (25%) believed the grading standard was generally too lenient, which is interesting considering that this is the high-stake cut-off point which has the greatest consequence for the individual student.

Concerning marking, the results of the survey showed that it was most common for the teachers to assess the speaking test performances *on their own* without peer marking (42%) (See Table 2 below). However, a fair number (36%) reported that they assessed *some of the performances* in collaboration with colleagues, or that they used co-rating for *many* or *all of the performances* (19%). In general, teachers were positive towards co-rating and thought it would contribute to a more reliable and fair assessment; however, lack of time and heavy workload were the main reasons for this not taking place. Many teachers voiced concerns about this, as exemplified below:

> I experience that the oral part of the national tests is the part that risks being the least equal in terms of assessment. The written part is peer marked to a greater extent, but the oral part is usually left to the individual teacher. [Resp. 125]

Teacher comments also revealed that conditions varied at the participating schools. Some schools organised peer marking, whereas at other schools the teachers had to find time for this on their own.

Table 2. Frequency of item 24, responding to the question: "How were the oral tests rated the last time you participated?" (*N* = 267)

Scoring practices	**Count**	**%**	**Valid %**
All performances co-rated	33	12.4	12.5
Many performances co-rated	17	6.4	6.4
Some performances co-rated	96	36.0	36.4
All performances rated by teacher alone	112	41.9	42.4
All performances rated by another teacher	3	1.1	1.1
Other	3	1.1	1.1
Total	264	98.9	100.0
Missing	3	1.1	
	267	100.0	

5.3 Available resources

Three binary items (Q7-9) were used to ask the respondents about the availability of *human resources*, in the form of support from the school management, and *material resources*, in the form of rooms/facilities and recording equipment. The responses indicate that a majority of the participating teachers (62%) thought that they did not receive enough support from their school leaders, which is surprising considering the fact that the responsibility for the implementation of the oral national tests is entrusted to the head teacher who should plan the organisation together with his/her staff at the local school (Swedish National Agency for Education, 2018). In terms of material resources, half of the teachers (50%) stated that there were enough rooms/facilities to carry out the national EFL speaking test at their school, whereas the other half claimed there were not. A large majority of the teachers stated that recording equipment was available and provided by the schools (76%).

In the teacher instructions, it is suggested that as one pair/group prepares, another takes the test, following a rota schedule. As pointed out by many teachers, the logistics of administering the speaking tests are complex since there are many steps to organise, which is why support from the school management and extra staff is needed. Additionally, many teachers stated that it was stressful to provide meaningful tasks for the rest of the class who were not taking the test. Further, teachers working at schools where there was a shortage of rooms remarked that this created a stressful situation. An open-ended comment serves to summarise the complex situation:

> The tests are time-consuming if there isn't extra time set aside for them, and there isn't. More support from the school management is needed in order to have reduced teaching time during the period when the national speaking tests are carried out. It would also be possible to arrange the oral parts centrally at schools. In that way, teachers wouldn't have to feel that it is an impossible equation to administer the oral tests when it is supposed to be done in

> combination with the regular ongoing teaching; there is no special time set aside, no special rooms allocated and not explicitly expressed what the rest of the students should work with meanwhile. Often, we have a sort of parallel responsibility for the rest of the students at the same time as we conduct the tests in small groups, this is not good. [Resp. 70]

5.4 Perceived practicality

The respondents were finally asked general questions about their views of practicality. When asked to indicate how they perceived the practical implementation of the test (Q10) on a seven-point scale from very problematic to not at all problematic, respondents, in general, used the middle point of the scale (M = 4.2, SD = 1.8), indicating a somewhat neutral standpoint. However, the large standard deviation points to great variation in teachers' opinions. In addition, teachers thought, overall, that the practical implementation of the speaking tests was too time-consuming (M = 3.6, SD = 1.9); once again, however, with variation in responses. With regard to the test instructions (Q11), the respondents believed they were generally clear (M = 6.0, SD = 1.2) and easy to follow (M = 6.1, SD = 1.1). However, as pointed out by some teachers in the open-ended comments, it was not always possible to adhere to the instructions in practice:

> For example, the part about pupils preparing individually, without aids. There aren't three group study rooms and a classroom available to carry out the test. This means I can't guarantee that pupils prepare individually without aids. I collect their mobile phones and computers when they get the preparation materials. However, anyone at the school can lend them a mobile or a computer without me seeing it. I'm busy with the group before them who are taking the test. [Resp. 171]

5.5 The influence of background variables

As has been shown, responses to the survey questions revealed some variation among teachers' practices and views. To explore whether three background variables, gender, teaching experience and school size, could be related to these differences, tests of association were conducted.

To start with, potential gender differences were examined through a series of Pearson chi-square tests for the categorical variables (Q2*a-d*, 7-9 and 24), and Mann-Whitney U-tests for the ordinal variables (Q 6*a-b*, 10*a-b*, 11*a-b*, 17*a-d*, 20*a-c*). Results indicated that, in relation to Q11*a*, female teachers were somewhat more prone to perceiving the instructions for the national speaking tests as clear than the male teachers, U (260) = 5041.500, Z = -1.986, N = 260, p = .047. No other statistically significant relationships were found. Thus, in this study no clear gender differences with respect to how teachers administered, assessed and viewed practical aspects of the national EFL speaking tests were shown.

Next, Spearman correlation analyses were used to investigate the relationship between teaching experience and ordinal dependent variables. Teaching experience correlated positively with two items, Q6*a* and 6*b* (See Table 1, Appendix E), indicating that the more teaching experience the respondent had, the more likely it was that he/she intervened in the conversation if a student did not get enough speaking opportunity or if the students got stuck in the conversation. Pearson chi-

square tests were employed to assess the association between teaching experience and categorical variables. No significant associations were found. In general, then, teachers' responses differed only to a limited extent based on teaching experience.

The possible contribution of school size in relation to responses was investigated, measured by two self-reported background variables. Spearman rank order correlations, although weak, showed that the number of English teacher colleagues correlated positively with one item and negatively with three (See Table 2, Appendix E). In addition, the teacher-reported number of students per school correlated positively with one item and negatively with two (See Table 3, Appendix E).

The findings suggest that teachers at smaller schools experienced more practical problems with the speaking tests and found them to be more time-consuming than teachers at larger schools, possibly related to the fact that at smaller schools the implementation of the oral tests is left to the individual teacher to a greater extent. In addition, there seemed to be a positive correlation between school size and Q20*a*, revealing that, in general, teachers at larger schools believed the grading of the sample performances for the lowest pass level, grade E, was too lenient. This may imply a local 'assessment culture' (Inbar-Lourie, 2008) at larger schools.

There was also an indication that respondents working at schools with fewer English teachers perceived they had more support for scoring from the supplementary assessment materials provided by the national Agency of Education, intended to explain and exemplify the national performance standards, than teachers working at larger schools. This might, once again, point to a local assessment culture, where having access to many English teacher colleagues provides the opportunity to engage in teacher co-operation in grading to a greater extent, whereas having fewer colleagues leads to a higher reliance on the assessment materials for interpreting the performance standards.

6 Discussion and implications

The main aim of the study reported here was to explore teachers' implementation practices and views of practicality with regard to the speaking component of a high-stakes national test of EFL in the Swedish upper secondary school, thus highlighting *contextual* (Weir, 2005) and *consequential* aspects (Messick, 1989) of test validity.

The results revealed that there were variations in administration and scoring practices. This is obviously an issue that needs to be treated from different angles, including possible consequences for students. On the one hand, it could be argued that "this flexibility in practices (...) compromises the reliability of the NEST [the National English Speaking Test] as a standardized test" (Sundqvist et al., 2017, p. 18). On the other hand, as Bachman and Palmer (2010) emphasise, the context of a test is complex: "Not only may differing stakeholder groups have different values, but in many contexts assessments are subject to a variety of different laws and regulations. These often operate at different levels (e.g., school, district, state, nation), and are sometimes in conflict with each other and with societal or educational values" (p. 257). The national EFL tests are centrally designed and

developed, following rigorous rules of standardisation (Erickson & Åberg-Bengtsson, 2012). However, in accordance with political decisions, the responsibility for the implementation of the oral national tests is entrusted to the head teacher who should plan the organisation together with his/her staff at the local school level. The National Agency for Education (2018) therefore concludes that "[t]he most suitable organisation of the oral national tests may look different at different schools". As can be seen, the national tests are embedded in a decentralised school system, requiring local decisions to be made and local responsibility to be taken. An important question raised in the current study is therefore how far a centrally designed paired speaking test can be standardised in terms of uniform administration procedures when carried out in a decentralised school setting.

Bachman and Palmer (1996) note that "practicality is a matter of the extent to which the demands of the particular test specifications can be met within the limits of existing resources" (p. 36). In light of this, some critical aspects of the administration and scoring procedures of the national speaking tests will be discussed. First, the number of students per group is an issue in need of further attention. In the present study, the majority of teachers reported using pairs, although groups of three, and sometimes even four, were also common. As previous research suggests that group size may have an impact on test interactions (Nakatsuhara, 2011), a stricter regulation about the number of students to include per group may be advisable. However, and not to be overseen, the issue as such, namely possible effects of two or three test takers in the assessment of oral interaction, needs to be further researched to find out to what extent the number has, or does not have, a significant effect on results.

Furthermore, analyses of the open-ended comments revealed that a careful matching of students was an essential task to teachers. It is interesting to note that there were no explicit instructions in the test guidelines regarding this. In the speaking tests for the lower secondary levels, such advice exists. Considering the potential effect of interlocutor characteristics (e.g., O'Sullivan, 2002), it may therefore be recommended to include a similar passage regarding matching of students in the guidelines for the upper secondary school. However, here as well, more research is needed regarding possible effects of different principles to form constellations, for example investigating effects of conscious versus random matching of pairs.

Another important question concerns recordings, which has many advantages as it makes re-listening and co-rating possible. In this regard, it should be noted that as many as 49% of the teachers in the present study reported that they recorded the oral national tests. Results from the annual surveys carried out by the test constructors point to increasing use of recordings; in 2010, on average 30% of teachers at the upper secondary school level responded that they recorded the tests, whereas in 2017, as many as around 70% claimed to do so (National Assessment Project, 2017). Making recordings mandatory is highly desirable as it would enable systematic documentation of the oral tests. However, in an educational large-scale

context, practical implications concerning the feasibility of collecting and storing recorded documentation, as well as implications in relation to laws about personal integrity and data security first need to be thoroughly analysed.

As regards scoring practices, findings from this study and from the annual surveys conducted by the test developers, indicate that it was common for the teachers to assess the oral national tests of English on their own, even though co-rating is strongly recommended in the test instructions as a measure to increase inter-rater reliability. Considering this, it is still encouraging to see that as many as 55% of the teachers in the current study reported that they co-rated some, many or all of the performances. Whereas many teachers expressed positive attitudes towards co-rating, lack of time and heavy workload were the main reasons for this not taking place, pointing to a need of more resources. The fact that recording is not mandatory and practices thus vary complicates the issue further.

In an investigation of the Swedish school system carried out by the OECD (Nusche et al., 2011), it is argued that 'it is vital' to increase the reliability of the teacher-rated national tests. The authors of the report suggest external moderation, teacher training and professional development as possible measures to take:

> External moderation is essential to ensure consistency, comparability and equity of the teacher-based assessments. There are several options of doing this, such as employing a second grader (a teacher in the same subject) in addition to the students' own teachers, employing professionals for systematic external grading and/or moderation, or introducing a checking procedure by a competent authority or examination board. In any of these options, high quality training for all graders is essential to ensure professional assessment competencies. (p. 11)

This should be considered in relation to current on-going activities at the national level, where *external rating*, carried out by a teacher other than the student's own, preferably from another school unit, and a form of *co-rating*, whereby two teachers, one of whom holds the main responsibility, independently mark the tests are being tried out in a pilot project coordinated by the National Agency for Education (Swedish Ministry of Education and Research, 2017). However, as pointed out in the OECD reports, there is also a need for high quality training "to help teachers understand and interpret the grade criteria and moderate assessment judgments"(OECD, 2015, p. 156). As a complement to a more formalised organisation of co-rating, it is therefore motivated to invest further resources in assessment literacy training as part of teachers' professional development (Malone, 2017; Xerri & Vella Briffa, 2018).

When looking at teachers' responses to the questions concerning available resources, as many as 62% of the teachers in the present survey stated that they did not receive enough support from the school management. The National Agency for Education (2018) stipulates that it is the responsibility of the head teacher to organise the implementation of the national oral tests together with his/her staff, "so that they benefit students and teachers in the best way possible". The results thus indicate that the decentralised responsibility of the implementation of the oral

tests is not fully working in this regard. Furthermore, half of the respondents stated that there was a lack of rooms at their school, which, in addition to making the test administration stressful for teachers, made it difficult to organise the preparation time, which is to be carried out individually and privately. Given the potential inequality this may lead to between schools, it may be advisable to reconsider how, and perhaps even if, the 15-minute preparation time should be part of the test. Also, as previous research has revealed inconclusive results concerning the effects of pre-planning time on performance in paired peaking tests (Elder & Iwashita, 2005; Lam, 2015; Nitta & Nakatsuhara, 2014; Wigglesworth & Elder, 2010), this issue requires further attention.

Finally, it was shown that school size accounted for some of the variation in teachers' responses. It was suggested that teachers at smaller schools, with fewer English teacher colleagues, experienced more practical problems with the speaking tests and found them more time-consuming than teachers at larger schools. This could possibly be explained by the fact that at smaller schools, the implementation of the oral tests is left to the individual teacher to a greater extent than at larger schools, once again highlighting the responsibility placed on the school management to organise the oral national tests.

To sum up, two main implications can be discerned. Firstly, the oral national tests are an important concern for the whole school and the implementation should not be left to be solved by the individual teachers who are conducting the tests. Clearer routines and administrative support are needed. Secondly, the aspect of co-rating needs considerable attention, from conceptual as well as practical points of view. Here, as well, the responsibility should not be passed on to individual teachers but must be a common concern at the local school or municipal level.

7 Limitations and conclusions

There are two main limitations to this study. First of all, considering the fact that the school management has the overall responsibility for the implementation of the oral national tests in the decentralised Swedish school setting, a more comprehensive investigation would include their perspectives as well. This is an important avenue for future research.

Another potential limitation is sample representativeness. Although steps were taken to select 150 schools randomly, and the response rate at the school level was high (79%), the teachers who responded to the survey were self-selected. This could lead to bias, as self-selected respondents may have a special interest in, or opinion on the survey topic. However, since the findings of the present investigation are largely supported by data from the annual surveys conducted by the test constructors, it may be concluded that the respondents are fairly representative of the target population.

As mentioned in the introduction, speaking was the least tested skill among national language tests in the European school context (European Commission/EACEA/Eurydice, 2015), most likely related to the complexity of the administration and the high costs involved. The testing of complex language skills,

such as speaking, should not be avoided in high-stakes contexts, especially as teachers and students consider the testing of oral proficiency important (Bellhouse, 2018; Zimina, 2018). Instead, such tests should be systematically developed and validated. As Roever (2004) points out, with reference to Bachman and Palmer (1996) and Ebel (1964), we need to consider the consequences of not administering the test:

> Bachman and Palmer (1996) describe the practicality of a test as the ratio between the resources available and the resources needed. Simply put, the more expensive a test is, the less practical it is. In reality, the less practical a test is, the less likely it is to be used. While practicality has often been treated as the ugly stepchild of validity, it is in fact directly related to considerations of the consequences of test use. As Ebel (1964) points out in evaluating the consequential validity of a test, we also have to take into account the consequences of *not* administering the test. (p. 285)

Finally, it needs to be pointed out that national assessment in Sweden is in a process of considerable change, not least following a decision to digitalise the system within a few years' time. This will undoubtedly affect the assessment of oral language competence in several ways. In this, input from different stakeholders, among which teachers is an important group, seems an essential aspect of the development of valid and quality-assured products and procedures.

References

Ahlin, Åsa & Eva Mörk (2008), "Effects of decentralization on school resources", *Economics of Education Review, 27*(3):276–284.

Amjadian, Mohiadin & Saman Ebadi (2011), "Variationist perspective on the role of social variables of gender and familiarity in L2 learners' oral interviews", *Theory and Practice in Language Studies, 1*(6):722–728.

Bachman, Lyle F. & Adrian S. Palmer (1996), *Language testing in practice*. Oxford: Oxford University Press.

Bachman, Lyle F. & Adrian S. Palmer (2010), *Language assessment in practice*. Oxford: Oxford University Press.

Bellhouse, Gemma L. (2018), "Are teachers given sufficient tools as examiners in high-stakes language testing? A study of the new foreign language speaking component of the French Baccalauréat", in Xerri, Daniel & Patricia Vella Briffa (eds.), *Teacher involvement in high-stakes language testing*. Cham: Springer International Publishing, 85–103.

Berry, Vivien (1993), "Personality characteristics as a potential source of language test bias", in Huhta, Ari, Kari Sajavaara & Sauli Takala (eds.), *Language testing: New openings*. Jyvaskyla, Finland: Institute for Educational Research, 115–124.

Berry, Vivien (2007), *Personality differences and oral test performance*. Frankfurt: Peter Lang.

Black, Paul, Christine Harrison, Jeremy Hodgen, Bethan Marshall & Natasha Serret (2011), "Can teachers' summative assessments produce dependable results and

also enhance classroom learning?", *Assessment in Education: Principles, Policy & Practice, 18*(4):451–469.

Brooks, Lindsay (2009), "Interacting in pairs in a test of oral proficiency: Co-constructing a better performance", *Language Testing, 26*(3):341–366.

Brown, Annie (2003), "Interviewer variation and the co-construction of speaking proficiency", *Language Testing, 20*(1):1–25.

Brown, James Dean (2018), "Standardized and proficiency testing", in Liontas, John I. (ed.), *The TESOL Encyclopedia of English Language Teaching*, 1–8.

Cheng, Liying, Yoshinori Watanabe & Andy Curtis (eds.) (2004), *Washback in language testing: Research contexts and methods*. London: Lawrence Erlbaum.

Council of Europe (2001), *Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*. Cambridge: Cambridge University Press.

Csépes, Ildikó (2009), *Measuring oral proficiency through paired-task performance*. Frankfurt am Main: Peter Lang.

Davis, Larry (2009), "The influence of interlocutor proficiency in a paired oral assessment", *Language Testing, 26*(3):367–396.

East, Martin (2015), "Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform", *Language Testing, 32*(1): 101–120.

Ebel, Robert L. (1964), "The social consequences of educational testing", in *Proceedings of the 1963 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 75–81.

Elder, Catherine & Noriko Iwashita (2005), "Planning for test performance: What difference does it make?", in Ellis, Rod (ed.), *Planning and task performance in a second language*. Amsterdam: John Benjamins, 219–238.

Erickson, Gudrun (2017, Nov.), *National assessment of foreign languages in Sweden*. Retrieved from https://nafs.gu.se/digitalAssets/1671/1671355_national_assessm_of_foreign_lang_in_sweden2017.pdf

Erickson, Gudrun & Lisbeth Åberg-Bengtsson (2012), "A Collaborative approach to national test development", in Tsagari, Dina & Ildikó Csépes (eds.), *Collaboration in Language Testing and Assessment*. Frankfurt am Main: Peter Lang, 93–108.

European Commission/EACEA/Eurydice (2009), *National testing of pupils in Europe: Objectives, organisation and use of results*. Brussels: Education, Audiovisual and Culture Executive Agency (EACEA P9 Eurydice).

European Commission/EACEA/Eurydice (2015), *Languages in secondary education: An overview of national tests in Europe – 2014/15. Eurydice Report*. Luxembourg: Publications Office of the European Union.

ffrench, Angela (1999), *Study of qualitative differences between CPE individual and paired test formats (Internal UCLES EFL report)*. Cambridge: University of Cambridge Local Examinations Syndicate.

Fulcher, Glenn (2003), *Testing second language speaking*. London [u.a.]: Longman.

Gustafsson, Jan-Eric & Gudrun Erickson (2013), "To trust or not to trust? – teacher marking versus external marking of national tests", *Educational Assessment, Evaluation and Accountability, 25*(1):69–87.

Harlen, Wynne (2005), "Teachers' summative practices and assessment for learning – tensions and synergies", *Curriculum Journal, 16*(2):207–223.

Harris, Douglas N. & Tim R. Sass (2011), "Teacher training, teacher quality and student achievement", *Journal of Public Economics, 95*(7):798–812.

Highlights report: languages in secondary education (2015, September 29). Retrieved from http://www.european-net.org/2015/09/eurydice-report-languages-in-secondary-education/

Holmström, Christian (2018, March 23), *Friskolor i Sverige* [Independent schools in Sweden]. Retrieved from https://www.ekonomifakta.se/Fakta/Valfarden-i-privat-regi/Skolan-i-privat-regi/Antal-friskolor-i-Sverige/

IBM Corp. (2017), IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Inbar-Lourie, Ofra (2008), "Language assessment culture", in Shohamy, Elana & Nancy Hornberger (eds.), *Encyclopedia of Language and Education* (Vol. 7) (2nd ed.). New York: Springer, 285–300.

Iwashita, Noriko (2001), "The effect of learner proficiency on interactional moves and modified output in nonnative–nonnative interaction in Japanese as a foreign language", *System, 29*(2):267–287.

Jin, Yan (2018), "Practicality", in Liontas, John I. (ed.), *The TESOL Encyclopedia of English Language Teaching*, 1–6.

Jones, Brett D. (2007), "The unintended outcomes of high-stakes testing", *Journal of Applied School Psychology, 23*(2):65–86.

Kormos, Judit (1999), "Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams", *Language Testing, 16*(2):163–188.

Lam, Daniel M. K. (2015), *Assessing interactional competence: The case of school-based speaking assessment in Hong Kong*. Unpublished doctoral thesis. Edingburgh: University of Edingburgh, UK.

Lazaraton, Anne (2002), *A qualitative approach to the validation of oral language tests*. Cambridge: UCLES/Cambridge University Press.

Malone, Margaret E. (2017), "Training in language assessment", in Shohamy, Elana, Iair G. Or & Stephen May (eds.), *Language Testing and Assessment*. Cham: Springer International Publishing, 225–239.

Messick, Samuel A. (1989), "Validity", in Linn, Robert L. (ed.), *Educational Measurement*. New York: Macmillan, 13–103.

Messick, Samuel A. (1996), "Validity and washback in language testing", *Language Testing, 13*(3):241–256.

Moss, Pamela A. (1998), "The role of consequences in validity theory", *Educational Measurement: Issues and Practice, 17*(2):6–12.

Nakatsuhara, Fumiyo (2006), "The impact of proficiency level on conversational styles in paired speaking tests", *Cambridge ESOL Research Notes*, 25:15–20. Retrieved from: https://www.cambridgeenglish.org/images/23144-research-notes-25.pdf

Nakatsuhara, Fumiyo (2009), *Conversational styles in group oral tests: How is the conversation co-constructed?*. Unpublished doctoral thesis. Essex: University of Essex, UK.

Nakatsuhara, Fumiyo (2011), "Effects of test-taker characteristics and the number of participants in group oral tests", *Language Testing, 28*(4):483–508.

National Assessment Project (2017), *Kursproven i engelska för gymnasieskolan* [National tests of English for upper secondary school]. Retrieved from https://nafs.gu.se/prov_engelska/engelska_gymn/resultat

Nitta, Ryo & Fumiyo Nakatsuhara (2014), "A multifaceted approach to investigating pre-task planning effects on paired oral test performance", *Language Testing, 31*(2):147–175.

Norris, John M. (2008), *Validity evaluation in language assessment*. Frankfurt: Peter Lang.

Norton, Julie (2005), "The paired format in the Cambridge Speaking Tests", *ELT Journal, 59*(4):287–297.

Nusche, Deborah, Gábor Halász, Janet Looney, Paulo Santiago & Claire Shewbridge (2011), *OECD Reviews of Evaluation and Assessment in Education: Sweden*. Paris: OECD.

Ockey, Gary J. (2009), "The effects of group members' personalities on a test taker's L2 group oral discussion test scores", *Language Testing, 26*(2):161–186.

OECD (2015), *Improving Schools in Sweden: An OECD Perspective*. Paris: OECD.

O'Loughlin, Kieran (2002), "The impact of gender in oral proficiency testing", *Language Testing, 19*(2):169–192.

O'Sullivan, Barry (2000), "Exploring gender and oral proficiency interview performance", *System, 28*(3):373–386.

O'Sullivan, Barry (2002), "Learner acquaintanceship and oral proficiency test pair-task performance", *Language Testing, 19*(3):277–295.

O'Sullivan, Barry (2013), "Assessing speaking", in Kunnan, Anthony (ed.), *The companion to language assessment*. Hoboken, NJ: John Wiley & Sons, 156–171.

Popham, James W. (2009), "Assessment literacy for teachers: Faddish or fundamental?", *Theory Into Practice, 48*(1):4–11.

Reemann, Edith, Ene Alas & Suliko Liiv (2013), "Interviewer behaviour during oral proficiency interviews: A gender perspective", *Eesti Rakenduslingvistika Ühingu aastaraamat / Estonian Papers in Applied Linguistics, 9*:209–226.

Rice, Jennifer King (2010), *The impact of teacher experience. Examining the evidence and policy implications* (Brief No. 11). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Roever, Carsten (2004), "Difficulty and practicality in tests of interlanguage pragmatics", in Boxer, Diana & Andrew D. Cohen (eds.), *Studying speaking to inform second language learning*. Clevedon: Multilingual Matters, 283–301.

Ryan, Katherine (2002), "Assessment validation in the context of high-stakes assessment", *Educational Measurement: Issues and Practice, 21*(1):7–15.

Sundqvist, Pia, Peter Wikström, Erica Sandlund & Lina Nyroos (2017), "The teacher as examiner of L2 oral tests: A challenge to standardization", *Language Testing, 35*(2):217–238.

Swedish Ministry of Education and Research (2017), *Förordning (2017:1106) om en försöksverksamhet med datorbaserade nationella prov, extern bedömning och medbedömning* [Ordinance (2017:1106) regarding a pilot project involving digitalised national tests, external rating and co-assessment]. Stockholm: Swedish Ministry of Education and Research.

Swedish National Agency for Education (2011), *Kommentarmaterial till ämnet engelska.* [Subject commentary: English]. Retrieved from https://www.skolverket.se/undervisning/kommentarer/kommentarmaterial.

Swedish National Agency for Education (2016a), *English 5, Spring Term 2016. Lärarinformation – inklusive bedömningsanvisningar till Delprov A Focus: Speaking* [English 5, Spring Term 2016. Teacher information – including scoring guidelines for Subtest A Focus: Speaking]. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education (2016b), *Utvärdering av den nya betygsskalan samt kunskapskravens utformning (Dnr 2014:892)* [An evaluation of the new grading scale and the design of the knowledge requirements]. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education (2018), *Att organisera muntliga delprov* [Organising the oral national subtests]. Retrieved from https://www.skolverket.se/undervisning/gymnasieskolan/nationella-prov-i-gymnasieskolan/genomfora-och-bedoma-prov-i-gymnasieskolan

Swedish Schools Inspectorate (2017), *Bedömningsprocessernas betydelse för likvärdigheten – Ombedömning av nationella prov 2016* [The importance of the scoring procedures for equity – Remarkings of national tests 2016]. Retrieved from http://www.skolinspektionen.se > Publikationer.

Urdan, Timothy C. & Scott G. Paris (1994), "Teachers' perceptions of standardized achievement tests". *Educational Policy, 8*(2):137–156.

Weir, Cyril J. (2005), *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Wigglesworth, Gillian & Cathie Elder (2010), "An investigation of the effectiveness and validity of planning time in speaking test tasks", *Language Assessment Quarterly, 7*(1):1–24.

Winke, Paula (2011), "Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter", *TESOL Quarterly, 45*(4):628–660.

Winke, Paula, Susan Gass & Carol Myford (2013), "Raters' L2 background as a potential source of bias in rating oral performance", *Language Testing, 30*(2):231–252.

Wiswall, Matthew (2013), "The dynamics of teacher quality", *Journal of Public Economics, 100*:61–78.

Xerri, Daniel, & Patricia Vella Briffa (2018), "Introduction", in Xerri, Daniel & Patricia Vella Briffa (eds.), *Teacher involvement in high-stakes language testing*. Cham: Springer International Publishing, 1–7.

Zimina, Evgeniia Vitalievna (2018), "Teacher involvement in the Russian national exam in foreign languages: Experience and perspectives", in Xerri, Daniel & Patricia Vella Briffa (eds.), *Teacher involvement in high-stakes language testing*. Cham: Springer International Publishing, 245–261.

Appendix A. Variables included in the present study

Items and response options

2. *How were the national speaking tests administered the last time you participated?*

a) *time period*

successively during the test period/ more concentrated during one or a couple of days/ both

b) *timing*

during lessons/ outside lessons/ both

c) *recording*

with recording/ without recording/ both

d) *grouping of students*

in pairs/ in groups/ both

If the oral tests were carried out in groups (as opposed to in pairs), please indicate how many students were included per group.

---

6. I intervene in the conversation if…

a) a student doesn't get enough speaking space

b) the students 'get stuck'

never or almost never / seldom / sometimes / often / always or almost always

If you can think of other reasons for teacher intervention or would like to provide additional comments, please write them here:

---

7. The school has enough rooms/facilities to administer the national speaking tests in an efficient way.

Yes/ No

---

8. Recording equipment is available at the school.

Yes/ No

---

9. *The teachers at my school receive support from the school management to organize and implement the national speaking tests.*

Yes/ No

10. How do you perceive the practical implementation of the national speaking tests?

a) very problematic 1 – 2 – 3 – 4 – 5 – 6 – 7 not at all problematic

b) far too time-consuming 1 – 2 – 3 – 4 – 5 – 6 – 7 reasonably time-consuming

11. How do you perceive the instructions for the national speaking tests?

a) not at all clear 1 – 2 – 3 – 4 – 5 – 6 – 7 very clear

b) difficult to follow 1 – 2 – 3 – 4 – 5 – 6 – 7 easy to follow

14. *If you have any additional general comments on Section 1 of the survey: "The practical implementation of the national speaking tests, please write them here:*

17. *To what degree do you believe you have support from the following assessment materials when scoring the national speaking tests?*

a) generic assessment factors

b) benchmarked and commented samples of oral performances

c) national performance standards

d) additional assessment materials provided by National Agency for Education

minimal support 1 – 2 – 3 – 4 – 5 – 6 – 7 very large support

20. To what extent do you believe the grading of the recorded and commented samples of oral performances generally is…

a) *Grade E:* too strict/ adequate/ too lenient

b) *Grade C:* too strict/ adequate/ too lenient

c) *Grade A:* too strict/ adequate/ too lenient

24. How were the oral tests rated the last time you participated?

a) all performances were co-rated

b) many performances were co-rated

c) some performances were co-rated

d) all performances were rated by the teacher alone

e) other (box provided for comments)

---

*27. If you have any additional general comments on Section 2 of the survey: "*Assessment in relation to national regulatory documents and purposes of the test"*, please write them here:*

Appendix B. Frequencies of administration practices

Table 1. Frequencies of different aspects of the administration of the national EFL speaking tests (*N* = 267)

2*a*. Time period	**Count**	**%**	**Valid %**
successively	105	39.3	39.5
during a couple of days	130	48.7	48.9
both	31	11.6	11.6
Total	266	99.3	100.0
Missing	1	0.4	
	267	100.0	
2*b*. Timing			
during lessons	162	60.7	61.1
outside lesson	32	12.0	12.1
both	71	26.6	26.8
Total	265	99.3	100.0
Missing	2	0.7	
	267	100.0	
2*c*. Recording			
with recording	128	47.9	48.7
without recording	105	39.3	39.9
both	30	11.2	11.4
Total	263	98.5	100.0
Missing	4	1.5	
	267	100.0	
2*d*. Grouping of students			
in pairs	109	40.8	41.9
in groups	74	27.7	28.5
both	77	28.8	29.6
Total	260	97.4	100.0
Missing	7	2.6	
	267	100.0	

Appendix C: Assessment factors provided in teacher guidelines for the national test for course English 6 in Swedish upper secondary school

CONTENT

- intelligibility and clarity
- complexity and variation
    - different examples and perspectives
- coherence and cohesion, structure
- adaptation to purpose, recipient, situation and genre

LANGUAGE AND ABILITY TO EXPRESS ONESELF

- communicative strategies
    - to develop and advance the conversation
    - to solve linguistic problems, e.g., through reformulations, explanations and clarifications
- fluency and ease of speaking
- range, variation, complexity, clarity and accuracy
    - vocabulary, phraseology and idiomaticity
    - pronunciation and intonation
    - grammatical structures
- adaptation to purpose, recipient, situation and genre

Translated from Swedish

Appendix D. Performance standards for courses English 5 and 6

(Searchable from: www.skolverket.se/laroplaner-amnen-och-kurser)

English 5: Approximate pass level (Grade E): Common European Framework of Reference for Languages (CEFR) Level B1.2 (Council of Europe, 2001)

Grade E	Grade C	Grade A
In oral and written communications of various genres, students can express themselves in **relatively** varied ways, **relatively** clearly and **relatively** coherently. Students can express themselves with **some** fluency and **to some extent adapted** to purpose, recipient and situation. Students work on and make improvements to their own communications. In oral and written interaction in various, and more formal contexts, students can express themselves clearly and with **some** fluency and **some** adaptation to purpose, recipient and situation. In addition, students can choose and use **essentially** functional strategies which **to some extent** solve problems and improve their interaction.	In oral and written communications of various genres, students can express themselves in a way that is **relatively** varied, clear, coherent and **relatively** structured. Students can also express themselves with fluency and **some adaptation** to purpose, recipient and situation. Students work on and make **well grounded** improvements to their own communications. In oral and written interaction in various, and more formal contexts, students can express themselves clearly with fluency, and with **some** adaptation to purpose, recipient and situation. In addition, students can choose and use functional strategies to solve problems and improve their interaction.	In oral and written communications of various genres, students can express themselves in ways that are varied, clear, coherent **and structured**. Students can also express themselves with fluency and **some adaptation** to purpose, recipient and situation. Students work on and make **well grounded and balanced** improvements to their own communications. In oral and written interaction in various, and more formal contexts, students express themselves clearly, **relative freely** and with fluency, and also with adaptation to purpose, recipient and situation. In addition, students can choose and use **well** functioning strategies to solve problems and improve their interaction, **and take it forward in a constructive way**.

Appendix D. (Continued)

English 6: Approximate pass level (Grade E): Common European Framework of Reference for Languages (CEFR) Level B2.1 (Council of Europe, 2001)

Grade E	Grade C	Grade A
In oral and written communications of various genres, students can express themselves in a way that is **relatively** varied, clear, and **relatively** structured. Students can also express themselves with fluency and **some** adaptation to purpose, recipient and situation. Students work on and make **simple** improvements to their own communications. In oral and written interaction in various, and more formal contexts, students can express themselves clearly and with some fluency and **some** adaptation to purpose, recipient and situation. In addition, students can choose and use **essentially** functional strategies which **to some extent** solve problems and improve their interaction.	In oral and written communications of various genres, students can express themselves in a way that is relatively varied, clear, coherent and relatively structured. Students can also express themselves with fluency and **some** adaptation to purpose, recipient and situation. Students work on and make **well grounded** improvements to their own communications. In oral and written interaction in various, and more formal and complex contexts, students can express themselves clearly, **relative freely** and with fluency, and with adaptation to purpose, recipient and situation. In addition, students can choose and use functional strategies to solve problems and improve their interaction.	In oral and written communications of various genres, students can express themselves in ways that are varied, **balanced**, clear and structured. Students can also express themselves with fluency and adaptation to purpose, recipient and situation. Students work on and make **well grounded and balanced** improvements to their own communications. In oral and written interaction in various, and more formal and complex contexts, students can express themselves clearly, **freely** and with fluency, and with adaptation to purpose, recipient and situation. In addition, students can choose and use **well** functioning strategies to solve problems and improve their interaction, and **take it forward in a constructive way**.

Source: The Swedish National Agency for Education (2011)

Appendix E. Association between background variables and response variables

Table 1. Significant Spearman correlations between teaching experience and items 6*a* and 6*b* relating to teacher intervention during the test occasion (*N* = 267)

Survey item	Correlation with teaching experience		Valid *N*
6. I intervene in the conversation if…			
a. a student doesn't get enough speaking opportunity	r_s = .187	ρ = .003	252
b. the students 'get stuck'	r_s = .248	ρ = .000	249

Table 2. Significant Spearman correlations between self-reported number of English teacher colleagues and items 10*a*, 17*a*, 17*d* and 20*a* (*N* = 267)

Survey item	Correlation with number of English teacher colleagues		Valid *N*
10. How do you perceive the practical implementation of the national speaking tests?			
a. very problematic – not at all problematic (7-point scale)	r_s = -.136	ρ = .028	260
17. ***To what degree do you believe you have support from the following assessment materials when scoring the national speaking tests?***			
a. analytic assessment factors minimal support– very large support (7-point scale)	r_s = -.125	ρ = .045	259
d. subject commentary material provided by National Agency for Education minimal support– very large support (7-point scale)	r_s = -.131	ρ = .037	257
20. To what extent do you believe the grading of the recorded and commented samples of oral performances generally is…?			
a. Grade E (too strict/adequate/too lenient)	r_s = .163	ρ = .009	257

Appendix E. (Continued)

Table 3. Significant Spearman correlations between self-reported number of students per school and items 10*a*, 10*b* and 20*a* (*N* = 267)

Survey item	**Correlation with number of students**		**Valid *N***
10. How do you perceive the practical implementation of the national speaking tests?			
a. very problematic – not at all problematic (7-point scale)	r_s = -.186	ρ = .003	255
b. far too time-consuming – reasonably time-consuming (7-point scale)	r_s = -.124	ρ = .048	255
20. To what extent do you believe the grading of the recorded and commented samples of oral performances generally is…			
a. Grade E too strict/adequate/too lenient (3-point scale)	r_s = .202	ρ = .001	252