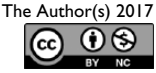
Article

LANGUAGE TESTING

The teacher as examiner of L2 oral tests: A challenge to standardization

Language Testing
2018, Vol. 35(2) 217–238



DOI: 10.1177/0265532217690782
journals.sagepub.com/home/ltj

Pia Sundqvist
Karlstad University, Sweden

Peter Wikström
Karlstad University, Sweden

Erica Sandlund
Karlstad University, Sweden

Lina Nyroos
Uppsala University, Sweden

Abstract

The present paper looks at the issue of standardization in L2 oral testing. Whereas external examiners are frequently used globally, some countries opt for test-takers' own teachers as examiners instead. In the present study, Sweden is used as a case in point, with a focus on the mandatory, high-stakes, summative, ninth-grade national test in English (speaking part). The national test has the typical characteristics of standardized tests and its main objective is to contribute to equity in assessment and grading on a national level. However, using teachers as examiners raises problems for standardization. The aim of this study is to examine teachers'/examiners' practices and views regarding four aspects of the speaking test – test-taker grouping, recording practices, the actual test occasion, and examiner participation in students' test interactions – and to discuss findings in relation to issues concerning the normativity and practical feasibility of standardization, taking the perspectives of test-takers, teachers/examiners, and test constructors into account. In order to answer research questions linked to these four aspects of L2 oral testing, self-report survey data from a random sample of teachers (N = 204) and teacher interviews (N = 11) were collected and quantitative data were analyzed using inferential statistics. Survey findings revealed that despite thorough instructions, teacher practices and views vary greatly across all aspects, which was further confirmed by interview data. Three background variables – teacher

Corresponding author:
Pia Sundqvist, Faculty of Arts and Social Sciences, Karlstad University, SE-65188 Karlstad, Sweden.
Email: pia.sundqvist@kau.se.

certification, work experience, gender – were investigated to see whether they could provide explanations. Whereas certification and gender did not contribute significantly to explaining the findings, work experience bore some relevance, but effect sizes were generally small. The study concludes that using teachers as examiners is a well-functioning procedure in terms of assessment *for* learning, but raises doubts regarding assessment *of* learning and standardization; a solution for test authorities could be to frame the test as non-standardized.

Keywords
Assessment, EFL, L2 interaction, oral tests, standardized testing, teacher practices

The present paper centers on challenges for students, teachers, examiners, and test constructors respectively, related to second/foreign language oral tests in general and the standardization of such tests in particular. Specifically, the focus is on teachers' practices and views as regards the speaking part of the mandatory national test in English used in the last year of secondary school in Sweden. This is a high-stakes, summative test, in which the test-taker's own teacher is the test administrator, instructor, and examiner (see "The national test in English in Sweden" section for more details about the test). All national tests used in Swedish schools are summative and share a twofold purpose: (1) to contribute to equity in assessment and grading and (2) to yield data for evaluation of goal-attainment (Swedish National Agency for Education, 2015). The examined English test is commonly perceived and described as a standardized test by important stakeholders, such as teachers/examiners and students/test-takers, evidenced in numerous discussions among members of the largest Facebook group for English teachers in the country (4500+ members; "English in grades 6–9") (see also Jönsson & Thornberg, 2014). In addition, several evaluations of the English national test show that it is greatly appreciated by teachers who, among other things, mention the fact that the national test assists them in interpreting the syllabus and in assigning final grades (Erickson & Börjesson, 2001; Naeslund, 2004; Velling Pedersen, 2013).

However, the Swedish Schools Inspectorate – an authority commissioned by the government to scrutinize schools – has recently raised concerns about its validity as a standardized test. For example, in a series of reports from a large-scale re-marking of the English national test (comprehension and writing parts) as well as national tests in other school subjects, the Inspectorate points to large discrepancies between national test grades assigned by the students' own teachers and test grades assigned by other teachers in the re-markings (Swedish Schools Inspectorate, 2011, 2012a, 2012b, 2013). The Inspectorate's findings were heavily publicized in the media and people in general, politicians, and the teachers themselves started to distrust teacher assessments (Gustafsson & Erickson, 2013). Although the test constructors have criticized the "suboptimal design" of the Inspectorate's research and defended the English national test as well as the use of teachers as examiners (Gustafsson & Erickson, 2013, p. 70), questions about the standardization of this test and the crucial role that teachers play remain.

As for high-stakes second/foreign language speaking tests, internationally it is uncommon to use the test-takers' own teacher as the examiner (cf. Roca-Varela & Palacios, 2013), but some countries adopt such a procedure. For example, Norway uses teachers as examiners in English, leaving a choice for teachers to assess speaking in a testing situation

or as part of classroom activities (Hasselgren, 2000). In New Zealand, an innovative large assessment reform known as *interact* has recently been implemented in foreign language education, and teachers are required to gather "three instances of interaction for summative grading purposes" over the school year (East, 2015, p. 5). Various resources have been made available to support the teachers in assessment. In his study, East (2015) raises an important issue connected to the standardization of oral tests. In essence, he brings up the tension between assessment *for* learning (formative assessment) as opposed to assessment *of* learning (summative assessment) (see also East, 2016). An important conclusion East draws about the assessment procedure of *interact* is that it may be trying to fulfill two potentially irreconcilable functions:

> On the one hand, using a series of genuinely authentic interactions as evidence of spoken proficiency is intuitively appealing, and setting up stand-alone assessments for *interact* focuses on the interaction as a *test* and potentially compromises the opportunity to collect evidence of genuine spontaneity. On the other hand, collecting lesson-embedded or "real life" evidence challenges fundamental notions of standardization and reliability that traditionally inform assessments that are used for high-stakes or accountability purposes. (East, 2015, pp. 115–116)

This tension also exists in Sweden, our key example on which this study is based. Like New Zealand, Sweden constitutes an interesting case in point because the Swedish testing system is atypical in combining a high level of faith in teachers (scoring their own students on a single test occasion) with the characteristics of a typical accountability system (Lundahl & Tveit, 2014). When compared, there are differences in operationalization between Sweden (snapshot assessment/one occasion) and New Zealand (ongoing assessment/several occasions).

In second/foreign language (L2) oral testing research, the topic of language teachers' practices and views is under-researched (cf. Roca-Varela & Palacios, 2013). The present paper seeks to address this deficit and to contribute to the discussion on standardization in L2 oral testing. We specifically discuss the findings in relation to questions concerning standardization norms and the practical feasibility of standardization, taking the perspectives of test-takers, teachers/examiners, and test constructors into account.

L2 speaking tests and standardization

All tests are consequential for learners. Assessing learner performance warrants the fulfillment of educational objectives and impacts, for example, grades and access to higher education, and high-stakes standardized tests are particularly consequential in this respect. As many scholars have pointed out, challenges for testing L2 oral proficiency include matters such as topic familiarity, interlocutor proficiency, test-taker relations, and task understandings (see, e.g., Brooks, 2009; Davis, 2009; Galaczi, 2008; Iwashita, 2001; Lazaraton & Davis, 2008; May, 2011). Oral tests present additional challenges to standardization, as social interaction is a joint achievement and external conditions are difficult to control; thus, conversations do not lend themselves well to standardization (Nyroos & Sandlund, 2014).

It has been argued that research on L2 oral proficiency is limited (Moeller & Theiler, 2014) and research on L2 oral tests even more so (Enright, 2004). Further, although language teachers "have unique insight into the collateral effects of tests" (Winke, 2011,

p. 633), their views are rarely included in evaluations of large-scale testing programs; this is surprising as teachers are "well positioned to recognize discrepancies between classroom and test practices" (p. 633) (cf. Johnson, 2013; Norris, 2008). Moreover, two recent research overviews of testing L2 speaking skills (Roca-Varela & Palacios, 2013; Sandlund, Sundqvist, & Nyroos, 2016) confirm that there are indeed few studies examining how teachers administer, carry out, assess, and grade L2 oral proficiency tests.

Globally, the most common format for high-stakes speaking tests is to use one examiner (who is a native speaker of the target language) together with one test-taker in an oral proficiency interview (OPI), even though paired and group tests are also used to a certain extent (Fulcher, 2003). Using teachers as examiners is much less common, but ever since the implementation of the ninth-grade national test in English in Sweden in 1998 (i.e., the particular test targeted in this study), teachers have been in charge of most of the testing process, that is, from introduction and preparation to carrying out the test and, ultimately, the assessment and grading of test-takers' performance (Erickson & Börjesson, 2001). This set-up means that the teacher has a threefold role as teacher, test administrator, and examiner during the test.

With regard to standardization, many scholars have attempted to define what constitutes a standardized test. Bachman (1990) mentions three characteristics. First, standardized tests are "based on a fixed, or standard content, which does not vary from one form of the test to another" (p. 74). This content may be based on a theory of language or on a specification of the test-takers' expected needs. Second, standard procedures for administering and scoring the test are used. Third, standardized tests are carefully tried out through empirical research, specifically with regard to their measurement properties and what measurement scale they provide. Further, Green (2014) describes standardized tests in a similar fashion. He argues that such tests are "built to tightly defined specifications, administered to all assessees under the same defined and controlled conditions, and scored according to clear rules" (Green, 2014, p. 241). Also Cizek (2012) emphasizes the importance of these specific, systematic, and uniform conditions for a standardized test, which are necessary to ensure that the test is employed for its intended purposes. By and large, these definitions are all in line with Bachman's (1990) definition of a standardized test (for more definitions of standardized tests and also criticism against standardized tests, see, e.g., Hughes, 1989; Moss, Pullin, Gee, Haertel, & Jones Young, 2008; Phelps, 2007; Zoghlami, 2014).

Regardless of which version of a standardized test is being used, equivalent test scores should indicate the same level of proficiency. Brown and Abeywickrama (2010) highlight that most secondary schools around the globe use standardized tests "to measure students' mastery of the standards or competencies that have been prescribed for specified grade levels" (p. 103), and these tests need to be fair, successfully administered, and perceived as meaningful by test-takers. Altogether, standardized tests should yield scores that are valid and reliable and consistent regardless of, for instance, the individual examiner (Bachman, 1990; Brown & Abeywickrama, 2010; Fulcher, 2003; Messick, 1989; Winke, 2011).

Bachman (1990) also describes norm-referenced and criterion-referenced tests. Whereas the former type of test is designed "to enable the test user to make 'normative' interpretations of test results" (p. 72), the latter type is designed "to enable the test user to interpret a test score with reference to a criterion level of ability or domain of content"

(p. 74). The national English test investigated here is a criterion-referenced test. With this discussion of L2 speaking tests and standardization as a backdrop, the next section focuses on teachers as examiners.

The teacher/examiner matters

The speaking test examined here is assessed by the test-takers' own teacher and as it is part of the high-stakes national test, the test score/grade is clearly important for the English subject final grade. In the Swedish context, it should be noted that only *certified* teachers are allowed to assign final grades for subjects and, therefore, it is appropriate to take a closer look at certification as a possibly relevant teacher characteristic in the present study. For instance, certified English teachers are expected to be well informed about standardized test procedures, as they must have completed a specific English teacher's degree program, which includes assessment training (www.skolverket.se), whereas uncertified teachers may not be as knowledgeable in this regard. Thus, in Sweden, the certification requirement may be seen as one mechanism for ensuring the standardization of assessment of the test examined in this study. Other background factors that may influence teachers' approaches to test procedures include teaching experience and academic qualifications.

One scholar who has researched teacher characteristics is Jacob (2012), who rightly argues that teachers are indeed central to learner achievement and that educational outcomes "depend more on the quality of the teacher a student is assigned to than on any other factor outside of the home" (p. 11). However, research reveals conflicting findings as regards the relations between *teacher experience*, *advanced degrees*, and *certification* on the one hand, and *student achievement* on the other. Jacob (2012) investigated evidence on all these variables and found small to null effects. Further, Kane, Rockoff, and Staiger (2007, 2008) compared the effectiveness of certification (comparing certified, uncertified, and alternatively certified teachers), using student data from New York City public schools (mathematics and reading comprehension tests), and although there were large differences in student achievement between teachers of the same status of certification, only small effects on learners' test performance were found between teachers across certification types. Additionally, whereas certification had little impact on test scores, work experience had a substantial effect. More specifically, teachers made "long strides in their first three years" (Kane et al., 2007, p. 64) but after that, there was little experience-related improvement. It is interesting to note that a second study on teacher effectiveness in New York presents almost identical findings (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2006). Another large-scale American study found that student achievement gains were systematically related to observable teacher (and school) characteristics, clearly showing that the teacher matters (Rivkin, Hanushek, & Kain, 2005). However, the effects were generally small. Moreover, as in the studies above, teachers' "learning curve appears to be quite steep in the first year or two of teaching before flattening out" (p. 435). Furthermore, they found "absolutely no evidence that having a master's degree improves teacher skills" (p. 449).

Scholars have also examined the effects of other variables. In studies on speaking tests, it has been found that gender as well as familiarity with the examiner may influence interaction (Amjadian & Ebadi, 2011) and assessment results (O'Sullivan,

2002). In our own work, we have seen that test-takers display divergent understandings of speaking test tasks and that certain task management strategies are rated less favorably than others, which influences assessment (Sandlund & Sundqvist, 2011). In addition, teachers'/examiners' understandings of how tasks should be handled have been shown to steer the interactional trajectory between test-takers in particular directions, thereby influencing performance (Sandlund & Sundqvist, 2013). Research also shows that the role of the interlocutor in paired/group tests is important for the reason that he or she could influence scores (Davis, 2009; Galaczi, 2008; Gan, 2010; Iwashita, 2001) as well as interaction (Lazaraton & Davis, 2008). Altogether, in light of the issue of standardization of speaking tests, it is evident that several factors influence test-taker performance and assessment, which is why we have included some of the frequently considered factors discussed in this section in the present study.

The national test in English in Sweden

In a recent paper on collaborative assessment and equity in assessment, influential Swedish assessment scholars (unconnected to the group of national test constructors) rightly establish that all national tests in Sweden are standardized tests (Jönsson & Thornberg, 2014). As for the national test in English, it clearly has Bachman's (1990) characteristics for standardized tests: (1) test content always builds on the core content stated in the curriculum; (2) all teachers/examiners are provided with detailed instructions regarding matters such as preparation, administration, and scoring – all of which they need to abide by; and (3) the test constructors try out test versions empirically in a rigorously monitored process (Erickson, 2009). Moreover, it is a proficiency test, which is intended to measure test-takers' global English proficiency by the use of three parts: one part tests speaking skills ('the National English Speaking Test', NEST); a second part tests receptive skills (listening and reading comprehension); and a third part tests writing skills (Erickson, 2012). At the time of the data collection, there was a window of 20 weeks in the spring for schools to administer the speaking test, whereas the other two had to be taken on specific dates (Swedish National Agency for Education, 2015). A previous study used different versions of the NEST and results showed that the grade mean for a group of learners improved significantly over two months (Sundqvist, 2009); thus, the actual date for when the NEST is offered can be consequential. Finally, over the years, the test constructors have conducted several test evaluations in which teacher satisfaction with the test has been assessed. In general, these evaluations have yielded positive results, indicating that most English teachers in Sweden are happy with how the test works (e.g., Naeslund, 2004; Velling Pedersen, 2013).

Scores on the NEST are interpreted in relation to criteria for grade levels A–E described in the curriculum; if criteria are not met, the grade is F (Swedish National Agency for Education, 2011a). The criteria, in turn, are to a large extent aligned with the communicative abilities described in the *Common European Framework of Reference for Languages* (Council of Europe, 2001). A passing grade (E) corresponds to level B1.1 (Council of Europe, 2001; Swedish National Agency for Education, 2011b).

Teachers are provided with a 31-page instruction booklet (eight pages deal directly with the NEST) and a CD with sample test recordings for different grade levels, but examiner scripts are not included. Teachers may prompt students if they run into difficulties, but as a

general principle, the teacher should remain in the background of students' conversations (Swedish National Agency for Education, 2013). The booklet provides comments on the recordings as well as information about the awarded grades and relevant criteria (National Assessment Project, 2015). The use of audio recordings is strongly recommended because they make re-listenings and collaborative assessment possible. However, recordings are not required. Despite annual recommendations from the test constructors, the proportion of teachers/examiners who record the test decreased from 41% in 1998 to 22% in 2007 (Velling Pedersen, 2007). According to Erickson (personal communication), the percentage of recordings has long remained stable at 20–25%. The faith put in teachers' professionalism by the National Agency for Education and test constructors is thus strong, and for the sake of stakeholders, not least test-takers, it is crucial for the system to work. However, as recordings are not required, systematic random collections of test recordings cannot be made, leaving little evidence of teachers' actual practices during the NEST. Potentially, this undermines standardization, making this test a suitable case in point for probing the issue of how standardization is carried out by teachers in an operational test setting.

Research questions

Bearing in mind the core characteristics of standardized tests, this study uses the English national speaking test in Sweden to address the challenges regarding and arguments about standardization that are outlined above. The present study is organized in terms of five research questions. Research questions (RQs) 1–4 concern teachers'/examiners' practices in carrying out the NEST, specifically regarding *test-taker grouping, recording practices, the actual test occasion*, and *participation in test interactions*. These four aspects of how teachers/examiners carry out the test provide information as to whether the NEST is conducted in a standardized way (see the "L2 speaking tests and standardization" section). Each aspect is regarded as relevant to the normativity and practical feasibility of standardization, and the results are discussed for each aspect in relation to arguments about standardization.

RQ1 How are test-takers grouped together in the speaking test?

RQ2 Are audio recordings used? If so, for what purpose(s)?

RQ3 What characterizes the actual test occasion as regards *time* and *place*?

RQ4 Do teachers/examiners participate in test interactions?

RQ5 concerns some background variables (teacher characteristics; see "The teacher/examiner matters" section), on the hypothesis that they might explain any differences found.

RQ5 To what extent do certification, work experience, and gender explain possible differences found in teachers'/examiners' practices/behavior/views?

Survey and interview data are used to answer RQs 1–4; survey data alone are used for RQ5.

The present study

This study adopts a mixed-methods research design. Data were drawn from a web-based survey among teachers as well as from teacher interviews, and both data sets were collected as part of a large project, Testing Talk (Swedish Research Council, reg. no. 2012-4129). In the project, one aim was to collect questionnaires from at least 200 secondary school English teachers.

Participants

In total, our sample consists of 204 teachers (see Table 1), 157 women and 47 men, from 25 to 66 years of age ($M = 45$; $SD = 9.5$). All but six reported having a teachers' degree. As regards a state-issued certification, 164 (80%) were certified English secondary school English teachers and 40 (20%) were not.

The teachers were selected through a two-step stratified random sampling procedure. First, the sample was stratified based on a division of municipalities into 10 types established by the Swedish Association of Local Authorities and Regions project (SALAR), such that each municipality type (SALAR category) was proportionally represented (http://skl.se/tjanster/englishpages.411.html). This was followed by a selection of schools for each SALAR category (every fifth school from an alphabetical list for each category), using the SIRIS database (http://siris.skolverket.se/siris/f?p=Siris:1:0) provided by the National Agency. After that, email addresses to the contact person(s) listed at school homepages were retrieved manually and the survey could be sent out (whenever multiple names were listed, all were contacted). The email instruction was to forward the survey to all English teachers at the school. In a first round, the survey was sent to 109 schools. It was open for four weeks (reminder after two). This resulted in 195 responses, making a second round necessary (another 10 schools; open three weeks; reminder after two). After round two, the target number had been reached (>200 responses). This sampling procedure was intended to allow for inferences about the general population of English teachers throughout Sweden.

A separate group of teachers (who participated in the Testing Talk project but were not questionnaire respondents) were interviewed ($N = 11$, Teacher 1–Teacher 11, all women). All interviewees had long work experience. They worked at four schools, two situated in a sparsely populated municipality and two in a large city. These 11 English teachers were all interviewed individually at the local school where they were employed during the spring semester of 2014.

Data

Survey data. Quantitative data from the survey were analyzed using inferential statistics (see the "Data analysis" section below). The purpose of the questionnaire was to yield self-report data about teachers' experiences of and views on the NEST and to learn about their assessment and grading practices. The survey included several items that elicited teachers' opinions about matters relating to standardization, which were suitable to use for the purpose of this study. All these items/questions are presented in Appendix A. The

Table 1. Number and percentage of English teacher respondents per SALAR category

SALAR category[a]	*N*	%
Metropolitan municipalities	23	11.3
Suburban municipalities	39	19.1
Large cities	61	29.9
Suburban municipalities to large cities	12	5.9
Commuter municipalities	13	6.4
Tourism and travel industry municipalities	5	2.5
Manufacturing municipalities	13	6.4
Sparsely populated municipalities	4	2.0
Municipalities in densely populated regions	20	9.8
Municipalities in sparsely populated regions	14	6.9
Sweden	204	100.0

[a]**S**wedish **A**ssociation of **L**ocal **A**uthorities and **R**egions project; official translations.

survey was constructed by the research team and built loosely on methodology brought forward in Horwitz's (1987) research on beliefs about language learning. Prior to its distribution, the survey went through extensive piloting, both in terms of content and technical procedures (Sundqvist, Sandlund, Nyroos, & Wikström, 2013).

The survey content was divided into three sections. The first focused on background information. For the purpose of this study, three background variables were examined: teacher certification (yes/no), work experience (less than 5 years/6–10 years/11–20 years/more than 20 years), and gender. Of these variables, certification and work experience are relatively strongly associated (χ^2 38.298, $df = 7$, $p = .000$; $\varphi_c = .433$). The second section tapped into teachers' practices as regards carrying out the test. The third covered the test occasion, assessment, and grading (see Appendix A). For this study, we focus on a subset of survey questions about teacher practices that are relevant to the standardization of NEST and background variables that previous research has shown may play a role.

Interview data. To supplement the quantitative findings, qualitative data from one open-ended survey question and data from teacher interviews were used. The 11 teachers were interviewed by either Researcher One or Researcher Three. A semi-structured format was adopted (Dörnyei, 2007). The interview guide had four parts: (1) background information; (2) preparing for and conducting the speaking test; (3) assessing and grading the test; and (4) speaking tests in general. All interviews were audio recorded and transcribed. The shortest interview was 29 minutes and 35 seconds, the longest 52:02 (mean: 45:30). As with the survey data, we were particularly interested in information from the participants that had to do with their views on the NEST and how they described the process of "carrying out NEST standardization" in practice.

Test material. The test material, including everything from the booklet to the actual test (described above), was also collected. In the analysis, the test material is referred to in order to frame the survey and interview responses.

Data analysis

All statistical tests were run in IBM SPSS Statistics 22. An alpha level of .05 was set; exact *p*-values are reported throughout. Pearson's chi-squared (χ^2) and Cramér's Phi (φ_c) were used for tests of association between nominal variables (e.g., the association between teachers' work experience and their grouping practices for the NEST). In correlation analyses (linear regression) involving ordinal data, Spearman's rank order correlation coefficient (r_s) was used. For example, Spearman was used to correlate teachers' work experience with how difficult they thought it was to grade the test ("I find it difficult to grade the speaking test"). The Mann–Whitney U test (Mann & Whitney, 1947) was used for associations between a dichotomous nominal independent variable (e.g., gender) and an ordinal dependent variable (e.g., Likert scale survey responses, such as "I feel uncomfortable when there is silence during the test"). To calculate effect size, the Mann–Whitney U tests were complemented with the rank-biserial effect size (Cureton, 1956; Kerby, 2014). The rank-biserial coefficient (r_{RB}) ranges from 0 to +/−1. The effect sizes used in this study may be roughly interpreted along the lines of Cohen's conventions: .2 is a small effect size, .5 medium, and .8 large (Aron, Aron, & Coups, 2005).

Results

Practices and views affecting standardization were revealed in the data. In presenting the results, findings based on survey data are presented first, followed by what the interviews revealed. The results are presented in the same order as the research questions.

Grouping test-takers

This section focuses on results for RQ1, which has to do with how the teachers grouped their students for the NEST. With regard to grouping test-takers, the booklet for the 2013 test (i.e., the last test prior to our survey) instructed teachers to arrange "a conversation between two or more students" (p. 9); a slightly different phrasing was used later: "[t]wo or possibly three students perform the task [of taking the test] together" (p. 17) (Swedish National Agency for Education, 2013). Further, the booklet advised teachers not to have students at very different proficiency levels in the same pair/group, as such combinations may affect performance negatively, and other interpersonally unfortunate groupings should also be avoided; for instance, students who generally do not get along well should not be in the same group.

The results of teacher responses to Question 13 ("How do you generally group your students for the national speaking test?") revealed that the majority of teachers (60.8%) used groups of three students, whereas dyads were used by 23.5% and groups of four or more by 15.7% (see Table 2). Further, in response to Question 14 ("Who decides which students belong to what group for the national speaking test?"), about half of the teachers decided which students to group together after consulting with the students (51.5%). Almost as many made the decision on their own (46.6%) and a few let the students decide (2.0%). When asked how important they think it is that students are at a similar proficiency level (Question 15), 66.2% responded "important" and 14.7% "very important". That is, the great majority followed the

Table 2. Frequencies for each response option, responding to the question "How do you generally group your students for the national speaking test?"

Response options (Q13)	*N*	%
Dyads	48	23.5
Groups of 3	124	60.8
Groups of 4 or more	32	15.7
Total	204	100.0

recommendation. Slightly less than a fifth did not ("somewhat important", 17.6%; "not very important", 1.5%), but that does not necessarily mean that these teachers considered a similar proficiency level among test-takers to be unimportant.

Our interview data revealed several underlying reasons for how students were paired/grouped for the test. One teacher made a link between the practice of recording and selecting group sizes: "Since we are recording now, we had them in pairs, since it's easier to have just two voices in the recording" (Teacher 2, interview). Another teacher argued that groups of three worked better as there was "more input into the conversation, it's more of a group conversation" (Teacher 3, interview). Another common reason for how students were grouped was the teacher's view of their proficiency level, whereas others grouped students purely on the basis of social relations.

In sum, most teachers treated the grouping of students seriously, as they believed this aspect matters for how comfortable students feel, which in turn may affect their performance. Notably, even in this small sample, the teachers gave varying reasons for *how* they grouped students. The survey and interview responses, thus, indicate variation in practices that may contribute to compromising the reliability of the NEST as a standardized test, specifically in terms of how consistently the test is administered.

Recording the test

With regard to RQ2, that is, whether audio recordings are used and for what purpose(s), as mentioned, there is a strong recommendation to record the NEST, but results showed that teacher practices differed greatly (see Table 3). About a quarter of all teachers claimed to record the test, but the majority did not.

Question 21 in the survey was a matrix of six items (*a–f*). Item *b* ("Teachers co-assessing the speaking test is common at my school") is particularly relevant in relation to recording practices, as co-assessment is typically done by the teachers sharing recordings with one another. The same holds for item *e* ("It happens that I wait a day or so before I grade a student's speaking test"), as teachers who use recordings may delay their assessment until after they have reviewed their recordings. Teacher responses to Question 21 are presented in Table 4. For a majority (72.3%), co-assessment was uncommon. This finding aligns with results for Items *a* and *e*, which revealed that it was common to grade test-takers' performance in conjunction with the test occasion rather than to wait a day or two. In addition, most teachers clearly appreciated the booklet and the CD (Item *d*: "Agree", 64.2%; "Strongly agree", 21.1%).

Table 3. Frequencies for each response option, responding to the question "How common is it that you record the speaking test?"

Response options (Q18)	*N*	%
Very uncommon (it happens rarely or never)	68	33.3
Uncommon (it happens occasionally)	62	30.4
Common (it happens often)	21	10.3
Very common (it is a habit)	53	26.0
Total	204	100.0

Table 4. Frequencies for each level of teachers' agreement with statements describing their assessment and grading of the NEST

Q21	Statement	*N* = 204	Strongly disagree	Disagree	Agree	Strongly agree
a	I grade student performances directly after the test.	*n*	10	37	91	66
		%	4.9	18.1	44.6	32.4
b	Teachers co-assessing the speaking test is common at my school.	*n*	80	68	36	20
		%	39.2	33.3	17.6	9.8
c	I find it difficult to grade the speaking test.	*n*	50	124	28	2
		%	24.5	60.8	13.7	1.0
d	I find great support for assessment and grading in the material (instructions and sample CD) provided by the National Agency for Education	*n*	5	25	131	43
		%	2.5	12.3	64.2	21.1
e	It happens that I wait a day or so before I grade a student's speaking test.	*n*	89	69	40	6
		%	43.6	33.8	19.2	2.9
f	I give individual feedback about the speaking test to my students.	*n*	3	27	98	76
		%	1.5	13.2	48.0	37.3

Analysis of responses to the open-ended Question 19 revealed varying reasons for recording/not recording the NEST. A majority of the responses dealt with arguments *against* recording, and the single most prevalent of these concerned lack of time for listening afterwards, as one written answer revealed: "The tests are enormously time-consuming to administer. It is just not possible for me to go back and listen to them again on top of that" (Respondent 35, Q19, survey). Others commented that the technical aspect of recording required extra work. A lack of recording equipment was also mentioned.

Thus, using recordings was considered pointless by many teachers, as the purpose would be to re-listen to the tests, something many felt they could not prioritize over other tasks. There was also a concern that students would feel less relaxed should they be recorded, and a belief that detailed notes made during the test (in combination with other

Table 5. Frequencies for each level of teachers' agreement with the statement "There are several (class)rooms available for me to organize and carry out the test."

Response options (Q17)	*N*	%
Strongly disagree	2	1.0
Disagree	30	14.7
Agree	62	30.4
Strongly agree	110	53.9
Total	204	100.0

oral assessment tasks during a school year) were enough to make an informed decision on the spot. However, among those who did not record, some had adopted their own solutions to manage fairness and equity in assessment: "We are always two teachers sitting in when students take the test – and then we do the assessment together. The workload is just too high to listen to recordings at a later point" (Respondent 122, Q19, survey).

Teachers who used audio recordings mentioned distractions in the face-to-face encounter during the test and argued that the recordings helped isolate the students' linguistic production. It is interesting to note that some did not appear to be familiar with the recommendation to record the tests, as is evident in this written comment: "Did not know that the National Agency recommended that" (Respondent 151, Q19, survey). Decisions to record were also brought up in the interviews in relation to teachers' preferences in terms of group sizes. One teacher mentioned that her school used to conduct the NEST in groups of four, and that assessment then became a problem "because we have seen that there is not enough time to assess them (.) properly" (Teacher 7, interview). She linked this problem to a relatively recent local decision to record the test and stated that "now it is easier because we have begun recording." Another teacher mentioned that she records her students because at times she had called on colleagues "for a second opinion" in order to avoid being too strict on her own students (Teacher 9, interview).

In sum, most commonly, *not recording* had to do with either time constraints or technical difficulties and *recording* with an idea that assessments would be more accurate and fair with additional listenings – and with teachers' wish to be able to focus solely on the test-takers in the actual test situation (forgetting about assessment right then and there). Altogether, the survey responses as well as the interviews reveal that the use of recordings and the reasons for recording or not tend to vary greatly. It seems clear that the recording practices must be more firmly regulated in the test materials, if this aspect of the NEST is to be viewed as standardized.

The time and place of the actual test session

This section focuses on the results for RQ3 concerning the *time* and *place* of the administration of the NEST. The instructions clearly say that it is up to the schools to make the local arrangements for the speaking test (Swedish National Agency for Education, 2013). A test normally lasts 15–25 minutes and rooms should be used in which the participants can be as undisturbed as possible.

With regard to the availability of (class)rooms (Question 17), almost a third (31.9%) of the teachers reported a lack of suitable rooms (see Table 5). Our interviews revealed that conditions at the participating schools also varied greatly: although some schools arranged particular speaking test days during which a substitute teacher worked with the rest of the class, others had to get the class started on independent work while they administered the test. Also, whereas one of the project schools had two days in February when all the tests were administered, another school administered tests "in between" other activities during the semester, so that all students completed their tests sometime between February and May. As mentioned above, this vast difference in terms of *when* tests are administered/taken may influence students' performance. One teacher stated that a major advantage of having a set speaking test day, in which students enter the test room, was that it "makes it more evident that it is a test situation, that now, we're doing the national test" (Teacher 2, interview). In contrast, at another school, a teacher utilized "empty slots" in her schedule for the tests, which meant that they were spread out over a longer period of time (Teacher 4, interview). It is evident that the local circumstances at particular schools (ranging from scheduling opportunities to the availability of rooms) put limits on the possibility of providing all test-takers with similar testing conditions.

Teacher participation in test interactions

Research question 5 asks whether the background factors teacher certification, work experience, and gender may explain some of the variation in teacher practices and views. Responses to four items in Question 20 (i.e., Items *b, c, d*, and *f*) relate to the possible participation by the teacher/examiner in test interactions. The results are presented in Table 6.

Clearly, most teachers reported staying silent. Further, about one in five said it was common to need to help students, and almost as many felt uncomfortable when silence arose. The instructions tell teachers to remain in the background, which some may interpret literally (hence the inclusion of Item *c*). Slightly more than half of all teachers claimed to sit away from their students. Overall, the interview data corroborated these findings, especially about being quiet and limiting participation in actual test interactions to providing assistance when problems arose, for example, when students/test-takers did not understand what to do. From the perspective of standardization, in comparison with the results for research questions 1, 2, and 3, the findings here are much less problematic because there appears to be greater consistency, even though there is still some variation in how the instructions are interpreted.

The influence of certification, work experience, and gender

As noted in "The teacher/examiner matters" section, teacher certification, work experience, and gender have been considered as potentially influential factors in test-taker performance and assessment in previous research. This possible influence is addressed by RQ5 and explored in this section.

The relationship between certification and testing practices was investigated by the use of independent samples Mann–Whitney U tests. The tests showed no significant group

Table 6. Frequencies for each level of teachers' agreement with statements describing how they would normally (inter)act in the actual test situation

Q20	Statement	*N* = 204	Strongly disagree	Disagree	Agree	Strongly agree
b	Usually I stay silent during the actual test and let the students manage the test on their own.	*n*	1	32	120	51
		%	.5	15.7	58.8	25.0
c	Usually I sit away from the students during the actual test.	*n*	31	58	58	47
		%	15.2	28.4	33.3	23.0
d	It is common that I need to help the students during the test.	*n*	20	138	42	4
		%	9.8	67.6	20.6	2.0
f	I feel uncomfortable when there is silence during the test.	*n*	86	80	36	2
		%	42.2	39.2	17.6	1.0

differences between certified and non-certified teachers in terms of how they responded to any of the survey questions that dealt with teachers' practices/views as regards carrying out the test, the actual test occasion, and assessment (more specifically Questions 15–18 and 20–21, ordinal data). At first, results for Question 21*f* (regarding giving individual feedback) indicated that non-certified teachers were somewhat more prone to giving individual feedback, $U\,(202) = 2672.000$, $Z = -1.987$, $p = .047$; $r_{RB} = .185$. However, on application of the Bonferroni correction for multiple comparisons, 21*f* was also non-significant. Pearson's chi-squared was used for tests of association between background variables and Questions 13 and 14 (both about grouping; nominal data). No significant associations were found. Thus, overall, teachers with or without certification had similar views about the test and they reported behaving in a similar way in the actual test situation as well as in subsequent work on assessment and grading, making it possible to conclude that certification is not a good predictor of teachers' practices regarding the NEST.

In order to examine the possible role of work experience in relation to teacher practices and views, Spearman correlation analyses were carried out. Work experience correlated positively with two variables and negatively with three (see Table 7). These correlations revealed that the more work experience, the more likely it was that teachers graded students immediately after the test and that they provided individual feedback. Further, there seemed to be a possible correlation between work experience and how difficult teachers thought it was to assess and grade the test, although this association was non-significant given the Bonferroni correction for multiple comparisons. As for the statistical tests involving the questions concerning grouping students, for work experience, a significant association was found with Question 14, that is, "Who decides which students belong to what group for the national speaking test?" (χ^2 19.953, $df = 6$, $p = .003$). The effect size was relatively small ($\varphi_c = .313$), but the tendency was that the more experienced teachers were more likely to let the students have a say in deciding how the groups should be composed.

With gender as a background variable, Mann–Whitney U tests initially seemed to reveal statistically significant differences for three of the ordinal variables, namely for

Table 7. Significant Spearman correlations between work experience and teachers' practices and views relating to administering and grading the NEST

Survey item	Correlation with work experience	
18. Do you usually record the test?	$r_s = -.218$,	$p = .002$
21*a*. I grade student performances directly after the test.	$r_s = .288$,	$p = .000$
21*c*. I find it difficult to grade the speaking test	$r_s = -.138$,	$p = .049$[a]
21e. It happens that I wait a day or so before I grade a student's speaking test.	$r_s = -.239$,	$p = .001$
21*f*. I give individual feedback about the speaking test to my students.	$r_s = .212$,	$p = .002$

[a]The Bonferroni-corrected alpha for this set of tests is .003; 21*c* is therefore not significant given a correction for multiple comparisons.

Items 20*b*, 20*c*, and 21*a*. For 20*b* ("Usually I stay silent during the actual test and let the students manage the test on their own"), the tendency was for female teachers to be slightly more prone to report that they stayed quiet during the test as compared to their male colleagues, U (202) = 2860.500, $Z = -2.649$, $p = .008$; $r_{RB} = .225$. Further, the tendency was very similar for 20*c* ("Usually I sit away from the students during the actual test"). Female teachers tended to be slightly more prone to respond that they seated themselves some distance away from the students, U (202) = 2988.500, $Z = -2.054$, $p = .040$; $r_{RB} = .190$. Regarding 21*a* ("I grade student performances directly after the test"), the pattern for the distribution of responses from male and female teachers was repeated in that it was more common for the women to grade test-takers' oral output directly, $U(202) = 3003.000$, $Z = -2.071$, $p = .038$; $r_{RB} = .186$. However, although these findings indicate that there could be gender-related differences in teacher practices, the Bonferroni-corrected alpha value for this set of tests ($\alpha = .003$) renders these differences non-significant.

In sum, the examination of the three background variables – certification, work experience, and gender – revealed that they explain some of the observed differences in teachers'/examiners' NEST practice, behavior, and views, but only to a limited extent.

Discussion and implications

As shown in several evaluations of the examined test (e.g., Naeslund, 2004; Velling Pedersen, 2013), teachers in Sweden generally appreciate the national English speaking test and its attached materials. However, the results of our study signal that interpreting and adhering to test instructions clearly presents a challenge to teachers as well as to standardization. There are several reasons for this. Our results reveal that local conditions have a great impact, such as whether teachers have access to recording equipment and quiet rooms for the NEST and whether substitute teachers are used. Regardless of the identified differences in terms of teacher practices, the interview data reveal that teachers want to ensure the best testing arrangements they can in accommodating the differing needs among their students.

Test arrangements include making decisions about the number of students per group, but the number of students is not always clearly stated in the instructions. Previous

research (e.g., Gan, 2010) has shown that group size may influence students' oral production in various ways. In our survey, groups of three students were most common, but fewer or more test-takers also occurred. In terms of group size, then, the results show that testing conditions are anything but equal. This is a threat to the validity of the examined test and, thus, a challenge for the test constructors.

Considering the fact that the test is perceived as standardized and that test materials look very similar year after year, it is possible to speculate that teachers skim (rather than carefully read) the booklet. As a consequence, important changes in the test instructions from previous years may be overlooked; this is an additional challenge for the test constructors.

With regard to using recordings, procedures also vary greatly, thereby not contributing to standardization in the sense of offering similar conditions in terms of possibilities for reassessment and collaborative assessment. Likewise, the extended time frame for the test is unfortunate from the perspective of test-takers, whose level of proficiency is likely to be lower in January as compared with May, making it an advantage to attend schools administering the test relatively late. In addition, evidenced in two data sets, our findings on teacher/examiner participation in test interactions – which may affect assessment and grading – also reveal different practices. Such practices present particular challenges for test-takers, teachers/examiners, and test constructors alike.

The findings presented here regarding certification and work experience are in line with previous research. Certification does not explain any of the identified differences in teacher practice, which may have to do with a higher propensity among uncertified teachers to leave the profession early as compared with certified teachers (cf. Kane et al., 2007). In contrast, work experience does explain some differences. For example, compared with colleagues who are newer to the job, experienced teachers more frequently grade the NEST in immediate conjunction with test administration. Moreover, they are more likely to provide individual feedback and consider students' opinions on how to compose test groups. The third background factor, gender, initially appeared to predict teacher responses to a greater extent than certification did. However, given a correction for multiple comparisons, gender was also found not to be an important explanatory factor in this study. Nevertheless, more research focusing on potential gender differences in test administration may be warranted.

On the whole, then, when the results of this study are *interpreted through a lens that assumes that standardization is necessary*, the examined test clearly does not measure up. The speaking test is undoubtedly part of a high-stakes assessment system, but it is not systematically operationalized as such. Thus, there is a tension between high-stakes accountability (which requires standardization, or rather, that the test is *perceived* as standardized by important stakeholders) and individual (context-specific) assessment. East (2015) identifies the same tension in his study from New Zealand, where he challenges the idea that standardization is possible, highlighting some problems associated with *interact*. For instance, *interact* is designed to embed ongoing speaking assessments within normal work in the classroom, but when perceived as part of a high-stakes assessment system, teachers find this impractical because there is a tendency to see each interactional episode as a test. In essence, the findings of the present study make an important contribution to the debate initiated by East (2015) as regards assessment *for* learning and assessment *of* learning.

In the Swedish case, the use of the teacher as examiner may be seen as a benefit in that teachers know their students/test-takers. Compared with using external examiners, teachers are in a better position to provide useful feedback and perhaps to know when particular students have the ability to resolve interactional problems on their own. In this regard, the use of teachers as examiners functions well. Moreover, although the NEST is a one-time test, it offers feed-forward opportunities; these, however, do not seem to be utilized in any systematic way. In brief, the teacher as examiner presents a problem with regard to standardization, but has other benefits.

Informed by the results of the present study, there are two possibilities for improving the NEST. First, the test could be kept unchanged but clearly and explicitly reframed as a non-standardized test, stressing its formative qualities. Second, if Swedish test authorities wish to maintain the NEST as a standardized test, the test and its operationalization could be redesigned in order to match key characteristics of standardized tests. Recommendations would be to state clearly, in rigorous test instructions, the number of test-takers per test occasion (preferably three) and to require/ensure the presence of two teachers/examiners who independently score the performances (alternatively one teacher/examiner, but with recordings as a requirement). Another recommendation would be the implementation of a script for the teacher/examiner to follow rather strictly, at least at the beginning of each test. Clearly, a choice needs to be made; this study shows that a half-measure does not work satisfactorily.

Conclusion

Despite the utility of the findings, the research had some limitations. For instance, we did not receive survey answers from all the schools that were approached. Most likely, the teachers at these schools never received our invitation to participate in the first place. Nevertheless, the teachers who did respond to the survey constitute a large sample of English teachers in Sweden. Another limitation is the relatively low number of interviews. However, the answers from the open-ended survey question and interviews were mutually reinforcing, and a combination of two data sets made triangulation possible, which is a strength of the current study. In addition, the teacher evidence collected and presented in the present study offers a rich stakeholder viewpoint.

Four aspects of the national English speaking test were investigated (test-taker grouping, recording practices, the actual test occasion, and teacher/examiner participation in test interactions), together with three background variables (certification, work experience, and gender), and discussed in relation to test standardization and the perspectives of test-takers, teachers/examiners, and test constructors. Overall, the study reveals many as well as major differences in teachers' practices and views regarding the test, where local conditions strongly influence practices. As has been argued, this flexibility in practices and the use of teachers as examiners, although it compromises the reliability of the NEST as a standardized test, may actually enable the NEST to function better both formatively as well as summatively. However, as long as equity in assessment and standardization remain important objectives of tests similar to the one examined here, the use of teachers as examiners creates conflicts.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Swedish Research Council (reg. no. 2012–4129).

References

Amjadian, M., & Ebadi, S. (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. doi:10.4304/tpls.1.6.722–728

Aron, A., Aron, E. N., & Coups, E. J. (2005). *Statistics for the behavioral and social sciences: A brief course* (3rd ed.). London: Prentice Hall International.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy*, *1*(2), 176–216. doi:10.1162/edfp.2006.1.2.176

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*(3), 341–366. doi:10.1177/0265532209104666

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment. Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.

Cizek, G. J. (2012). An introduction to contemporary standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–14). New York: Routledge.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika*, *21*(3), 287–290. doi:10.1007/BF02289138

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367–396. doi:10.1177/0265532209104667

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing*, *32*(1), 101–120. doi:10.1177/0265532214544393

East, M. (2016). *Assessing foreign language students' spoken proficiency: Stakeholder perspectives on assessment innovation*. Singapore: Springer.

Enright, M. K. (2004). Research issues in high-stakes communicative language testing: Reflections on TOEFL's new directions. *TESOL Quarterly*, *38*(1), 147–151. doi:10.2307/3588266

Erickson, G. (2009). Nationella prov i engelska – en studie av bedömarsamstämmighet. Retrieved from www.nafs.gu.se/digitalAssets/1319/1319572_bed.np-eng-g.erickson-2009.pdf

Erickson, G. (2012). National assessment of foreign languages in Sweden. Retrieved from www.ips.gu.se/om-ips/personal?userId=xericg

Erickson, G., & Börjesson, L. (2001). Bedömning av språkfärdighet i nationella prov och bedömningsmaterial. In R. Ferm & P. Malmberg (Eds.), *Språkboken* (pp. 255–269). Stockholm: Myndigheten för skolutveckling.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Education.

Galaczi, E. D. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89–119. doi:10.1080/15434300801934702

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, *27*(4), 585–602. doi:10.1177/0265532210364049

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Oxon: Routledge.

Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust? Teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability*, *25*(1), 69–87. doi: 10.1007/s11092–013–9158-x

Hasselgren, A. (2000). The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing*, *17*(2), 261–277. doi:10.1177/026553220001700209

Horwitz, E. K. (1987). Surveying student beliefs about language learning. In A. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 119–129). London: Prentice Hall.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Iwashita, N. (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. doi:10.1016/S0346–251X(01)00015-X

Jacob, A. (2012). Examining the relationship between student achievement and observable teacher characteristics: Implications for school leaders. *International Journal of Educational Leadership Preparation*. Retrieved from http://files.eric.ed.gov/fulltext/EJ997469.pdf

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education*, *28*(1), 91–105. doi:10.1080/02671522.2012.754229

Jönsson, A., & Thornberg, P. (2014). Samsyn eller samstämmighet? En diskussion om sambedömning som redskap för likvärdig bedömning i skolan. *Pedagogisk forskning i Sverige*, *19*(4–5), 386–402.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2007). Photo finish: Certification doesn't guarantee a winner. *Education Next*, *7*(1), 60–67.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631. doi:10.1016/j.econedurev.2007.05.005

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Innovative Teaching*, *3*(1), 1–9. doi:10.2466/11.IT.3.1

Lazaraton, A., & Davis, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly*, *4*(4), 313–335. doi:10.1080/15434300802457513

Lundahl, C., & Tveit, S. (2014). Att legitimera nationella prov i Sverige och i Norge – en fråga om profession och tradition. *Pedagogisk forskning i Sverige*, *19*(3–4), 297–323.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*(1), 50–60. doi:10.1214/aoms/1177730491

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127–145. doi:10.1080/154303.2011.565845

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education/Macmillan.

Moeller, A. J., & Theiler, J. (2014). Spoken Spanish language development at the high school level: A mixed-methods study. *Foreign Language Annals*, *47*(2), 210–240. doi:10.1111/flan.12085

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Jones Young, L. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. Cambridge: Cambridge University Press.

Naeslund, L. (2004). *Prövostenar i praktiken. Grundskolans nationella provsystem i ljuset av användares synpunkter*. Stockholm: Skolverket.

National Assessment Project. (2015). National Assessment Project. Retrieved from http://nafs.gu.se/english/?languageId=100001&disableRedirect=true&returnUrl=http%3A%2F%2Fnafs.gu.se%2F

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

Nyroos, L., & Sandlund, E. (2014). From paper to practice: Asking and responding to a standardized question item in performance appraisal interviews. *Pragmatics & Society*, *5*(2), 165–190. doi: 10.1075/ps.5.2.01nyr

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

Phelps, R. P. (2007). *Standardized testing primer*. New York: Peter Lang.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458. doi:10.1111/j.1468–0262.2005.00584.x

Roca-Varela, M. L., & Palacios, I. M. (2013). How are spoken skills assessed in proficiency tests of general English as a Foreign Language? A preliminary survey. *International Journal of English Studies*, *13*(2), 53–68. doi:10.6018/ijes.13.2.185901

Sandlund, E., & Sundqvist, P. (2011). Managing task-related trouble in L2 oral proficiency tests: Contrasting interaction data and rater assessment. *Novitas-ROYAL*, *5*(1), 91–120.

Sandlund, E., & Sundqvist, P. (2013). Diverging task orientations in L2 oral proficiency tests – a conversation analytic approach to participant understandings of pre-set discussion tasks. *Nordic Journal of Modern Language Methodology*, *2*(1), 1–21.

Sandlund, E., Sundqvist, P., & Nyroos, L. (2016). Testing L2 talk: A review of empirical studies on second language oral proficiency testing. *Language and Linguistics Compass*, *10*(1), 14–29. doi:10.1111/lnc3.12174/epdf

Sundqvist, P. (2009). *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary* (Dissertation). Karlstad University, Karlstad.

Sundqvist, P., Sandlund, E., Nyroos, L., & Wikström, P. (2013). Genomförande och bedömning av nationella muntliga prov i engelska: en pilotstudie. *KAPET*, *9*(1), 24–45.

Swedish National Agency for Education. (2011a). *Curriculum for the compulsory school, preschool class and leisure-time centre 2011*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2011b). *Kommentarmaterial till kursplanen i engelska*. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2013). *English. Ämnesprov, låsår 2012/2013. Lärarinformation inklusive bedömningsanvisningar till Delprov A. Årskurs* 9. Stockholm: Swedish National Agency for Education.

Swedish National Agency for Education. (2015). Nationella prov. Retrieved from www.skolverket.se/bedomning/nationella-prov

Swedish Schools Inspectorate. (2011). *Lika eller olika? Omrättning av nationella prov i grundskolan och gymnasieskolan. Redovisning av regeringsuppdrag [The same or different? The re-marking of national tests in compulsory school and upper secondary school. Report from a government commission]*. Stockholm: Swedish Schools Inspectorate.

Swedish Schools Inspectorate. (2012a). *Lika för alla? Omrättning av nationella prov i grundskolan och gymnasieskolan under tre år [The same for all? The re-marking of national tests in compulsory school and upper secondary school during three years]*. Stockholm: Swedish Schools Inspectorate.

Swedish Schools Inspectorate. (2012b). *Riktad tillsyn av bedömning och betygssättning hos skolor med stora avvikelser vid omrättning av nationella prov [Examination of assessment and grading at schools with great deviations in re-assessments of national tests]*. Stockholm: Swedish Schools Inspectorate.

Swedish Schools Inspectorate. (2013). *Olikheterna är för stora. Omrättning av nationella prov i grundskolan och gymnasieskolan, 2013 [The differences are too great. Re-assessing national tests in compulsory and upper secondary school, 2013]*. Stockholm: Swedish Schools Inspectorate.

Velling Pedersen, D. (2007). *Ämnesprovet 2007 i grundskolans årskurs 9. En resultatredovisning (Engelska)*. Stockholm: Skolverket.

Velling Pedersen, D. (2013). *Ämnesproven 2012 i grundskolans årskurs 9 och specialskolans årskurs 10 (Engelska)*. Stockholm: Skolverket.

Winke, P. (2011). Evaluation the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly*, *45*(4), 628–660. doi:10.5054/tq.2011.268063

Zoghlami, N. (2014). Testing L2 listening proficiency: Reviewing standardized tests within a competence-based framework. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 191–207). Bristol: Multilingual Matters.

Appendix A. Examined survey questions (sections 2–3)

	Question and response options (translated from Swedish)
13	*How do you generally group your students for the national speaking test?* Dyads / Groups of 3 / Groups of 4 or more
14	*Who decides which students belong to what group for the national speaking test?* Teacher decides / Teacher decides after consulting with the students / Students decide
15	*How important do you think it is that the students in a group are at a similar level of proficiency?* Not very important / Somewhat important / Important / Very important
16	*Together with the assessment guidelines, the National Agency for Education provides a CD with a number of sample tests. Do you usually listen to the CD?* I never listen to the CD / I rarely listen to the CD / I often listen to the CD, but not always / I always listen to the CD
17	*There are several (class)rooms available for me to organize and carry out the test.* Strongly disagree / Disagree / Agree / Strongly agree
18	*Do you usually record the speaking test?* Very uncommon (it happens rarely or never) / Uncommon (it happens occasionally) / Common (it happens often) / Very common (it is a habit)
19	*Although the National Agency for Education strongly recommends the use of recordings, very few teachers record their students (about 20%). Please comment briefly on why you choose to record your students or not.*
20	*Question 20 includes six statements (a–f), having to do with how you would normally act in the actual test situation. For each statement, tick the alternative that best describes your own practice. (Only Items b, c, d, and f reported here.)* *b. Usually I stay silent during the actual test and let the students manage the test on their own.* *c. Usually I sit away from the students during the actual test.* *d. It is common that I need to help the students during the test.* *f. I feel uncomfortable when there is silence during the test.* Strongly disagree / Disagree / Agree / Strongly agree
21	*Question 21 includes six statements (a–f) relating to the assessment and grading of the test. For each statement, tick the alternative that best describes your own practice.* *a. I grade student performances directly after the test.* *b. Teachers co-assessing the speaking test is common at my school.* *c. I find it difficult to grade the speaking test.* *d. I find great support for assessment and grading in the material (instructions and CD) provided by the National Agency for Education.* *e. It happens that I wait a day or so before I grade a student's speaking test.* *f. I give individual feedback about the speaking test to my students.* Strongly disagree / Disagree / Agree / Strongly agree