Interrater reliability in a national assessment of oral mathematical communication

Torulf Palm

Mathematical communication, oral and written, is generally regarded as an important aspect of mathematics and mathematics education. This implies that oral mathematical communication also should play a part in various kinds of assessments. But oral assessments of subject matter knowledge or communication abilities, in education and elsewhere, often display reliability problems, which render difficulties with their use. In mathematics education, research about the reliability of oral assessments is comparably uncommon and this lack of research is particularly striking when it comes to the assessment of mathematical communication abilities. This study analyses the interrater reliability of the assessment of oral mathematical communication in a Swedish national test for upper secondary level. The results show that the assessment does suffer from interrater reliability problems. In addition, the difficulties to assess this construct reliably do not seem to mainly come from the communication aspect in itself, but from insufficiencies in the model employed to assess the construct.

Communication of mathematics is generally regarded as an important part of mathematics and mathematics education. It is seen as a means to enhance learning of mathematics in general as well as an important mathematical competence in itself. It is included in many goal specifications such as the *American principles and standards for school mathematics* (NCTM, 2000), the framework for the international comparative study PISA (OECD, 1999) and the Danish KOM-project (Niss & Jensen, 2002). Both the oral and written modes of mathematical communication are also included in several of the Nordic countries' national curricula or syllabi documents (e.g. Danish Ministry of Education, 2007; Finnish

Torulf Palm
Umeå universitet

Palm T. (2008). Interrater reliability in a national assessment of oral mathematical communication. *Nordic Studies in Mathematics Education*, 13 (2), 49–70.

National Board of Education, 2004; Norwegian Directorate for Education and Training, 2007; Swedish National Agency of Education, 2001).

When oral mathematical communication is regarded as important and included in a curriculum there are several reasons for assessing this ability. One set of reasons pertains to the use of assessment for learning, the *formative function* of assessment (Wiliam, 2007). The information about the students' skills and understanding that is gathered from the assessment is used to guide the proceeding learning process. When assessments are used for helping students develop their oral mathematical communication abilities the advantage of the oral mode of assessment is obvious. But this mode of assessment may also be important for students' learning even when the assessment is not having a formative function, but are used for decisions about, for instance, grading. For example, when only written assessments are used students may form the belief that oral mathematical communication is not necessary in school mathematics and direct their learning activities accordingly. Also, since teachers' practice, in some aspects and under certain conditions, are influenced by externally mandated assessments (Barnes, Clarke & Stephens, 2000) such assessments can also more indirectly be important for the characteristics of classroom practice and thus student learning. For example, exclusion of oral assessments in externally mandated high-stakes tests might not constitute favourable conditions for such activities to be included in teaching and textbooks, which in turn would not be favourable conditions for students' to develop oral communication abilities. In addition to arguments of learning gains, the *alignment* (Webb, 1997) between curricular goals (such as oral mathematical communication) and an assessment system has several advantages when it comes to, for example, grading. One of them is that conclusions about students' attainment of curricular goals can be based on performances that are actual instances of the curricular goals and do not have to rely on, for example, correlation studies showing that the students' performances are relevant for judgements about whether particular learning goals have been attained.

Oral assessment can be defined as "assessment in which a student's response to the assessment task is verbal, in the sense of being 'expressed or conveyed by speech instead of writing'" (Oxford English dictionary) (Joughin, 1998, p. 367). This means that an assessment in which students' verbal responses are combined with other forms of responses, such as written solutions, still can be regarded as an oral assessment as long as the oral responses are assessed. A distinction can be made between two different qualities that can be assessed by oral assessment: (1) the students' communication abilities in themselves and (2) the subject matter knowledge that is demonstrated through the oral response (Joughin, 1998). The assessment of the latter quality has a long history within mathematics education, even if research about the quality of its use is scarce. The assessment of the communication abilities in themselves is a well-established part of language education and research within this area is vast, which is natural since language is (oral) communication in its very nature. The same does not seem to be valid for mathematics education and research in mathematics education. In fact, a search in the *Mathematics education database* (http://www.emis.de/MATH/DI.html) using the words 'communication', 'oral', and 'assessment', respectively, in the title did not render any English written research article that dealt with category 1, the assessment of mathematical communication abilities in themselves.

In several of the Nordic countries the national assessments and examinations include an oral part. For example, for reasons of alignment with the national steering documents some of the Swedish national tests in mathematics include such a part. For the more advanced courses in upper secondary level the oral part is designed to measure the mathematical communication abilities. However, the criterion of alignment between assessments and curricula documents is not enough to ensure that the assessments are high-quality assessments in which useful and proper interpretations from assessment scores can be made. As expressed by Kane et al. (1999, p. 6), "[i]f we are to have any confidence in a proposed interpretation, the evidence supporting the interpretation needs to substantially outweigh any evidence against the proposed interpretation". For this to happen we have to consider several criteria for the quality of educational assessments (Linn, Baker & Dunbar, 1991) or a broad interpretation of the concept of validity (Messick, 1989). However, "perhaps the most serious criticism of oral examinations concerns the level of reliability that are typically observed" (Raymond & Viswesvaran, 1991). Since reliability issues seem to be the weakest part in the argumentation for the usefulness of assessment scores from more direct assessments, such as oral assessments of mathematical communication, Kane et al. (1999) argue that for these assessments special attention should be given to reliability issues.

The aim of this study is to analyze the interrater reliability of an assessment model used for assessing oral mathematical communication in a Swedish national test for upper secondary level.

Reliability

To be able to make appropriate interpretations of assessment results the assessment would have to have the quality that the results could be replicated if the same students were assessed again under similar circumstances. Such consistency (or reproducibility) of assessment scores is called *reliability* (Crocker & Algina, 1986). That assessments possess this quality is far from evident.

> Whenever an examinee responds to a set of test items his or her score represents only a limited sample of behavior – responses to a subset of many possible items from a given domain obtained on one of many possible occasions. Consequently scores obtained under these conditions are fallible and subject to errors of measurement. Errors of measurement can be broadly categorized as random or systematic. Systematic measurement errors are those which consistently affect an individual's score because of some particular characteristic of the person or the test that has nothing to do with the construct being measured. [...] By contrast, random errors of measurement affect an individual's score because of purely chance happenings. (Crocker & Algina, 1986, p. 105)
>
> Both types of errors may cause test scores to be inaccurate and thus reduce their practical utility and thus are a source of concern in score interpretation. Random errors specifically reduce the consistency of the test scores and thus concern reliability.
>
> (Crocker & Algina, 1986, p. 106)

Nyström (2004) identifies three main areas threatening the reliability of educational assessments. First, since the tasks included in an assessment are only a subset of possible tasks in a domain the students' performance is dependent on the task sample. The interaction between tasks and persons has often been found to be substantial (see for example, Brennan & Johnson, 1995). In such cases generalizations of students' performances across tasks would be unreliable. The second area concerns temporal instability, which refers to the problem that the students' performances are assessed at one of many occasions and students' performances can differ from one occasion to another (often denoted test-retest reliability). Examples of circumstances that affect students' performances in this area are the quality of the last night's sleep and temporary personal problems. Thirdly, reliability is affected by interrater variation, which means that assessors can judge performances differently.

Interrater reliability

Interrater reliability can be seen as "the level of agreement between a particular set of judges on a particular instrument at a particular time. Thus, interrater reliability is a property of the testing situation, and not of the instrument itself" (Stemler, 2004, p. 2). Rater disagreement has different components. When raters are rating performances into ordered categories they may differ in the definition of the construct that is being measured or in the interpretation of the rating categories. Although in the past interrater reliability often has been seen as it was a single, unitary concept Stemler (2004) proposes that the statistical methods for computing interrater reliability most commonly reported in the literature can be classified into one of three categories: 1) consensus estimates, 2) consistency estimates, and 3) measurement estimates. "Each of these statistics will provide a statistical estimate of the extent to which two or more judges are applying their ratings in a manner that is predictable and replicable". (Stemler, 2004, p. 16). The estimates from each category have different assumptions, interpretations, advantages, and disadvantages. The choice of estimate to calculate in a study depends on the purpose at hand.

Assessing to which degree performances represent different levels of a construct, such as oral mathematical communication ability, will carry with it some subjectivity since the judgement of the performance will depend on the judges interpretation of the construct and the rating levels. Applying scoring rubrics is one way of imposing some objectivity into the assessment, which can be strengthened by also training judges how to interpret the scoring rubrics and apply the levels of the rating scale (Kane et al., 1999; Tierney & Simon, 2004).

Consensus estimates

The purpose of *consensus estimates* is to analyse the exact agreement among independent judges. These measures of interrater reliability will be high when observers come to exact agreement on how to apply the scoring rubrics. This indicates that they share a common interpretation of the construct and the scores they give can be treated as equivalent. Consensus estimates are often most useful when data are nominal (can be classified into categories) in nature and different levels of the rating scale represent qualitatively different ideas, or when data are ordinal (can be classified into categories that can be placed in order of precedence) in nature but different levels of the rating scale are assumed to represent a linear continuum of the construct, e.g. a Likert scale (Stemler, 2004). Examples of methods for calculating consensus estimates of interrater reliability are the computation of the percent-agreement statistic (Frick & Semmel, 1978) or Cohen's kappa statistic (Cohen, 1960, 1968).

Consistency estimates

Consistency estimates are based on the computation of the consistency of each judge's application of the scoring rubrics. For these estimates of interrater reliability to be high different judges need not make the same interpretation of the rating scale, and make the same classifications of performances, as long as they are consistent in their use of the scale when classifying the performances. Such consistency may sometimes be sufficient, for instance when scores can be corrected for differences in judges' severity. For example, if one rater consistently marks performances with one point more than other raters, then all the ratings by this rater can just be adjusted with one point. Consistency estimates are most useful when data are continuous in nature, although they can also be used when data are categorical if the categories of the rating scale represent an underlying continuum of a unidimensional construct (Stemler, 2004). Examples of methods for calculating the degree of consistency between judges are the computation of Pearson correlation coefficient or Spearman's rank coefficient (Glass & Hopkins, 1996) or Cronbach's alpha coefficient (Crocker & Algina, 1986).

Measurement estimates

Under the measurement approaches to interrater reliability, like for the consistency estimates, it is not necessary for raters to come to consensus on how to apply the scoring rubrics. The goal with this approach is often to estimate the severity of the judges and adjust students' scores accordingly. In this approach all available information from all judges are used (Stemler, 2004). Such information would include their ratings but may also include the judges' age, perceived item difficulty etc. "It is the accumulation of information, not the ratings themselves, that is decisive" (Linacre, 2002, p. 858). The factor analytic technique of principal component analysis (Harman, 1967) can be used to visualise how the use of more information makes the measurement estimates different from the consistency estimates. In the principal component analytic approach the amount of shared variance in the ratings that could be accounted for by the first principal component is determined. If this variance is high it indicates that the judges have reached agreement and are rating a common construct (Stemler, 2004). Other measurement estimates of interrater reliability can be computed through the use of generalizability (Shavelson & Webb, 1991), and through the use of the many-facets Rasch model (Linacre, 1994).

Oral national assessment in Sweden

The Swedish school system

The Swedish upper secondary school is governed by national steering documents. The syllabus for each mathematics course describes the mathematical knowledge to be taught and learned. The grading system is based on national criteria for four different levels of attainment of the content described in the syllabi. These grade levels are termed *Not passed*, *Pass*, *Pass with distinction* and *Pass with special distinction*. The national tests are criterion-referenced. Based on the results on the national test students receive a test grade determined by the in advanced decided cut scores. The teachers assign the final course grade, and this grade is based on both the national test result and other performances made during the course. The course grades from upper secondary school are used for admission to the university. Swedish upper secondary mathematics is divided in five consecutive courses, A–E. Course A is studied by all students and is often the only mathematics course taken by the students following a vocational program. The Social science students take at least courses A and B, and the Natural science students at least A–D.

Mathematical communication in the national mathematics syllabi

Communication, both oral and written, is emphasized in the 1994 Swedish upper secondary syllabi for mathematics as well as in the latest upper secondary syllabi that came into practice in the year 2000. Both sets of syllabi state that communication is one of ”four important aspects of the subject that permeate all teaching” (Swedish National Agency for Education, 1994, p. 47; Swedish National Agency for Education, 2001, p. 61), and that to ”present their thoughts orally and in writing” (Swedish National Agency for Education, 1994, p. 47; Swedish National Agency for Education, 2001, p. 60) is a goal to aim for. Consequently, the national grading criteria describe different levels of mathematical communication for different grades. In the 1994 syllabi (valid for students finishing their studies until 2002) the grading criteria below concern oral mathematical communication (Swedish National Agency for Education 1994, p. 49, author's translation). No criteria were given for the grade Pass with special distinction (the government's opinion was that the teachers would be able to define this level without the support of nationally defined criteria) (Swedish Ministry of Education and Research, 1993). In the syllabi that came into practice 2000 the grading criteria had been revised and also included descriptions of the highest grade Pass with special distinction.

Criteria for Pass
The student can with some support orally present the train of thought in their work with, and solution to, problems even if the mathematical language is not treated entirely correct.

Criteria for Pass with distinction
The student can orally with a clear train of thought present and explain the procedure in the problem solving in an acceptable mathematical language

Oral national tests of mathematical communication

In 1999 and from 2003 and onwards the Swedish upper secondary schools are offered to take part in an oral assessment as one part of the national test for the C-course. The main intention with this part is twofold; to support the teachers in assessing students' oral communication of mathematical ideas and thoughts and to support equivalent grading of such communication.

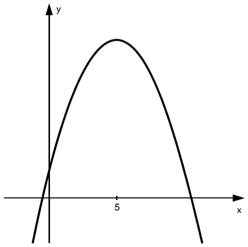
Swedish teachers are not provided extra time for carrying out the national tests. Therefore, it was seen as important to develop an assessment model for the oral part that requires as short teacher time as possible, while still maintaining sufficiently high reliability and validity. For this purpose the time frame for each student's oral performance was limited to 5 minutes (and 15 minutes to work with the task) in the 1999 assessment. Due to concerns about whether 5 minutes would be enough time to gather sufficient information for making reliable and valid ratings the available time for each student performance was increased to 10 minutes in the assessments used 2003 and onwards. However, these concerns were not based on research data.

In the assessment the students orally present their written work with one of the suggested tasks or another task chosen by their teacher. The judgement of the performance is based on the nationally administered scoring rubrics. The following is an example of a suggested task for the 1999 assessment (see task 2:2, author's translation).

To facilitate the focus on oral communication abilities it is suggested that the teacher monitors the students' work with the solution to the task and when needed interacts with the students so each student has a correct solution to present. The teacher and the students should have discussed the scoring rubrics before the actual assessment. In the first half of the actual assessment situation the students may present their work. During the other half the teacher or other students may ask questions about the presented content. This is intended to lead to a mathematical discussion where the assessed students have to engage in a multi-way discussion.

Task 2:2
The graph to a second-degree function is drawn in the coordinate system below.

a) Draw, in a coordinate system of your own, how the graph to the function's derivative may look like.
b) Draw, in a coordinate system of your own, how the graph to the function's second derivative may look like.



Your teacher will assess your work with respect to:
- How well you present the train of thought in your solution
- The mathematical language you use
- The mathematical knowledge you show

Since it was desirable that the teachers and students can choose the task to work with the oral test does not govern the tasks that are possible to choose. As a consequence the scoring rubrics are generic for all types of tasks. The trade-off for scoring rubrics with such universality is that they do not contain specific descriptors for each task solution the students are presenting (Tierney & Simon, 2004), which may make it more difficult for students and raters to interpret them. Such student and rater variability may be reduced by clarifying generic terms by also attaching task-specific exemplars (Wiggins, 1998). The assessment material supporting the teachers in the Swedish national assessments does include, in addition to scoring rubrics and experiences made in the tryouts that may facilitate the teachers' assessment work, written pieces of authentic student work and transcriptions of graded and commented oral presentations of this work. The support material for the 2003 assessment also includes audiofiles of these presentations. These audio recordings are available at the test institution's website (see address below).

In the 1999 oral assessment the teachers were to assess both the students' subject matter knowledge and their mathematical communication abilities. The communication abilities were divided into two aspects; Oral account of the train of thought and Mathematical language. The three aspects should be assessed separately. To attain the test grade Pass on this assessment the student has to show abilities corresponding to this level of performance on at least two of the subaspects. The requirement for the grade Pass with distinction is analogous to this rule. This shows the focus on mathematical communication since it is possible to acquire a certain test grade without performing up to this standard on the subaspect Subject matter knowledge. However, these aspects are not entirely independent of each other. A student could explain a procedure in a solution to a task that includes a common question with a clear train of thought and with a mathematical language with few deficiencies. However, it is likely that it is easier to make clear explanations with an appropriate mathematical terminology if the subject matter is well understood. Since there were no national criteria for the grade Pass with special distinction at this time no guidance were given for the requirements of this test grade. The scoring rubrics for the 1999 assessment are given in table 1.

Table 1. *Scoring rubrics for the 1999 oral national assessment for course* C

Assessment aims at	**Qualitative levels**	
	Pass	**Pass with distinction**
Oral account of the train of thought	Presents with some support the train of thought in the work with, and solution to, the task in such a way that the teacher and students understand.	Presents and explains the procedure in the task solving with a clear train of thought.
Mathematical language	The used mathematical language has considerable deficiencies but is understandable.	The student uses a mathematical language, when necessary, with only a few deficiencies.
Subject matter knowledge	The student solves, possibly with some support, a task that includes a simple and common question. Alternatively, the student solves parts of a more complex task.	The student solves a task that can be characterised by a question that is more difficult and/or uncommon. The student discusses and assesses when necessary his/her solution strategy.

Note. Author's translation.

The 2003 oral assessment only includes the assessment of mathematical communication abilities. The model for mathematical communication underlying this assessment is a further development of the model underlying the 1999 oral assessment. The support material states that the scoring rubrics, which are a further elaboration of the 1999 scoring rubrics, are based on the idea that the quality of an *oral presentation* of a mathematical solution or another mathematical issue is dependent on the three subaspects *existence and completeness of relevant descriptions and explanations*, the *structure* of the presentation, and the *mathematical terminology* used to convey the descriptions and explanations. The support material also includes a description of the background of the design of the assessment model and a more thorough description of the three subaspects of oral mathematical communication that the judgement should focus (for a more comprehensive presentation of the 2003 test material and the considerations underlying the choice of assessment model, see the test institution's website at http://www.umu.se/edmeas/np).

Method

Interrater estimate

In this study a consensus estimate will be used as a measure of the interrater reliability. This type of estimate is suitable for the study since (1) the credibility of the test grades assigned to students on the oral part of the national test is dependent on the extent the teachers can come to an exact agreement of how to apply the scoring rubrics, (2) in the authentic national test situation it is not possible to adjust students' test grades for teachers' severity in their judgements, and (3) the data in the form of the test grades the teachers' assign to students' performances are ordinal (can be classified into categories that can be placed in order of precedence) but the different levels of the grading scale are assumed to represent a linear continuum of the construct corresponding to each rated aspect, for example mathematical terminology use. There are a number of consensus estimates that can be used. One of the most popular estimates is Cohen's kappa statistic. It is intended to measure the level of agreement correcting for the proportion of times raters would agree by chance alone. However, this measure is controversial for several reasons, one being that it is dependent on the proportion of ratings falling into each rating category (Uebersax, 1987). In addition, in absence of an explicit model of raters' decision-making it has been argued that the statement that it is "a chance-corrected measure of agreement" is misleading (Uebersax, 2007). The consensus estimate chosen for this study is the percent-agreement statistic, The percent-agreement statistic is calculated by, for each pair of judges, adding the number of performances that are equally rated and dividing this number with the total number of performances rated by the two judges. A disadvantage with using this measure is the difficulty of comparing the results of the study with studies of assessment models with different numbers of rating categories. However, as a measure of the rating agreement between raters using this assessment model it is appropriate and the results can be directly compared with studies of other assessment models using four rating categories. In addition, the results are easy to interpret. But the statistic does not provide rich evidence about factors that may be the cause of disagreement. To collect data indicating whether the reason for disagreement is due to different definitions of the construct being measured or due to different interpretations of the category levels Svensson's method is used (Svensson, 1998). With this method it is possible to separately measure whether raters have different definitions of the scale levels or different definitions of the constructs being measured. Information about the first issue comes from the statistic *relative position,* RP. The value of RP can vary between -1 and 1, where a value of 0 would indicate that there is no systematic difference in the level of the ratings between two raters. A value of (-)1 would indicate a clear systematic difference in the level of ratings (the sign shows which rater has made the highest ratings).However, this measure does not take into consideration whether ratings from different raters differ in their concentration on the rating scale. The stastic the *relative concentration,* RC, expresses the extent to which the distribution of ratings from one rater is more centered to specific rating categories than the ratings from another rater. This statistic can also vary between -1 and 1 where the value 0 would indicate that there is no difference in two raters' concentration of their ratings on the scale. The *relative rangvariance,* RV, is a measure of the differences in the ratings that are not systematic to higher or lower levels of the scale or as a concentration of the ratings. This measure can vary between 0 and 1. When RV is 0 there are no such differences. A value of 1 would strongly indicate that the raters' definitions of the construct to be rated differ.

Procedure

The data for this study was collected in 1999 and the assessment situation in the study was designed to emulate the 1999 oral national assessment for course C. Since this assessment model includes the rating of both mathematical communication and subject-matter knowledge this provides the possibility to investigate the interrater reliability of the assessment of mathematical communication abilities in a national assessment as well as to compare this reliability with the interrater reliability of the assessment of students' subject matter knowledge that is demonstrated through the oral response. This is not possible in the 2003 assessment model. Furthermore, it makes it possible for future interrater reliability studies of mathematical communication, investigating other assessment models such as the Swedish 2003 national assessment model, to compare their results with the results based on the 1999 assessment model.

In the study a student presented his/her work according to the national test procedure outlined earlier in this paper. After the student's approximately 3-minute presentation of the solution to a task a 2-minute discussion of the work took place between the student and the author of this paper, who acted as the teacher in the study. In addition, ten external teachers were also present in the classroom during the act. Their role was to, independently, assess the student's performance in accordance with their interpretation of the scoring rubrics. Before this occasion they had studied the test material available to teachers, and the written work the student had made on the task. As in the authentic national test situation they rated the student's performance in relation to the three aspects Subject matter knowledge, Oral account of the train of thought and Mathematical language. They also assigned a total oral assessment test grade for the student performance according to the given procedure. Apart from following the national test procedure the ten raters also judged the certainty with which each rating was given. They could choose between the following four choices: (1) very uncertain, (2) somewhat uncertain, (3) somewhat certain, and (4) very certain. This procedure was repeated with another 5 students. The first three students presented different tasks and each of the following three students presented the same task as one of the first three students had presented. Thus, solutions to three different tasks were presented and two students presented solutions to each task.

Participants and tasks

The ten teachers (four women and six men) that participated in the study came from different parts of the country and participated in the development of the written parts of the national tests. They were of different ages but most of them were well experienced and highly engaged in their own teaching. Two of them had also been involved in writing textbooks. The six students participating in the study were recruited from the same upper secondary school in a middle-size university city. They were chosen by their teacher at the school and were chosen to represent different levels of mathematical performance. The tasks were assigned to the students from the set of tasks suggested in the national test material.

Results

The ten judges are combined to constitute 45 pairs of judges. Each pair of judges may agree in 1/6, 2/6, etc. of their ratings of the six student performances. The first column of table 2 displays the means of the percent-agreement values that are calculated for each of the 45 pairs of judges. The second and third columns display the lowest and highest percent-agreement values found among these 45 pairs of judges. For each judge we also calculated the mean of the percent-agreements for the pairs of judges including this judge. In other words, the agreements between judge A and B, A and C etc. are calculated and the mean of these values are computed. Then the same is done for each of the ten judges. The lowest and highest of these ten means are displayed in the fourth and fifth columns. The first row consists of values based on the judges' assigned test grades for all six students' performances. The second to the fourth row consists of values based on the judges' judgements of all of the six students' performances on the subaspects Oral account of the line of thought, Mathematical terminology, and Subject matter knowledge, respectively. The fifth to the tenth row consists of values based on the judges' judgements of the first three and the last three student performances on the three subaspects.

The results show that the interrater reliability of this oral assessment is not very high, and that the interrater reliability is not high for any of the three subaspects. When the students' performances were judged in relation to the two communication subaspects Oral account of the line of thought and Mathematical terminology the means of the percent-agreements between all pairs of judges over the six student performances were 55% and 51 % respectively. The corresponding mean for the subaspect Subject matter knowledge was 43 % and for the total oral assessment test grade 53 % (see table 2, column 1). For the two communication aspects the lowest percent-agreement between two judges is 17 % and the highest agreement is 83 %. The corresponding agreements for the subaspect Subject matter knowledge are 0% and 100% (table 2, column 2–3).

For each judge we also calculated the mean of the percent-agreements for the pairs of judges including this judge. In table 2 (columns 4–5) it can be seen that the lowest such mean is 33% for the subaspect Line of thought, 38% for Mathematical terminology, and 33% for subject matter knowledge. The highest means are 68%, 57% and 48%, respectively. Thus, although there are certainly differences between the judges in their ratings, no individual judge marks out as totally different in the judgements of the students' performances.

When comparing the ratings of the first three student performances with the ratings of the last three it can be seen that the mean of the

Table 2. *Agreements between all pairs of judges*

	Mean of all percent-agreements	Lowest percent-agreement	Highest percent-agreement	Lowest mean	Highest mean
				of percent-agreements linked to each individual judge	
Test grade	53	17	100	33	65
Line of thought	55	17	83	33	68
Mathematical terminology	51	17	83	38	57
Subject matter knowledge	43	0	100	33	48
Line of thought/ Students 1-3	61	0	100	37	74
Line of thought/ Students 4-6	49	0	100	22	59
Mathematical terminology/ Students 1-3	53	0	100	41	63
Mathematical terminology/ Students 4-6	49	0	100	37	59
Subject matter knowledge/ Students 1-3	44	0	100	30	52
Subject matter knowledge/ Students 4-6	43	0	100	37	52

percent-agreement statistic is lower for the last three students' performances on all subaspects, and especially for the subaspect Line of thought. However, the interrater reliability is still not very high even for the rating of the first three students' performances (see table 2, rows 5–10). The mean of the teachers' experienced certainty in their ratings is 2.7 on all of the three subaspects, which is very close to the value 2.5 that lies in the middle of the interval $1 \leq x \leq 4$, which includes the integers they could use in their certainty estimates.

The use of Swenson's method to gather information about the reasons for disagreemt provided data (see table 3 below) that indicates that the disagreement between raters seems to have been caused by both different definitions of the construct being measured (indicated by a high RV) and different interpretations of the rating categories (indicted by a high RP), but also on a difference in the centring of the ratings (indicated by a high RC). Table 3 shows that different definitions of the rating categories seems to be the main cause of disagreement when rating the subaspects Line of thought and Subject matter knowledge, while the centring of ratings seems to be a greater problem when rating the aspect Mathematical terminology.

Table 3. *Cause of disagreements between judges*

	Line of thought (*n* = 23)			Mathematical terminology (*n* = 29)			Subject matter knowledge (*n* = 37)		
	RP	RC	RV	RP	RC	RV	RP	RC	RV
$0 \leq x < 0.2$	30	70	70	62	41	79	41	54	76
$0.2 \leq x < 0.4$	22	26	9	28	28	14	16	22	5
$0.4 \leq x < 0.6$	35	4	18	3	24	7	24	14	14
$0.6 \leq x < 0.8$	13	0	0	7	7	0	19	11	3
$0.8 \leq x < 1$	0	0	4	0	0	0	0	0	3

Note. For each pair of rater distributions that do not coincide in more than three of the six ratings the statistics RP, RC and RV have been calculated. The table shows the proportions of those values that fall into each defined interval.

Discussion

Reliability is an important criterion for the quality of assessments. To be able to make appropriate interpretations of assessment scores the scores would have to show some consistency when students are assessed under similar circumstances. This study focuses on interrater variation, which is one of the major threats to reliability (Nyström, 2003). The study shows that, like many oral assessments of mathematical content knowledge, this Swedish national test model with a strong focus on oral mathematical communication abilities suffers from interrater reliability problems. The mean values of the percent-agreement estimates of 55% and 51% for the aspects Line of thought and Mathematical terminology, respectively, are not high especially taking into consideration that the judging of the students' performances only includes four grade levels. The line of thought and the mathematical terminology seem equally difficult to rate reliably. Since interrater reliability refers to a particular set of judges using a particular instrument at a particular time the results of interrater reliability studies are only valid for the particular assessment situation in which the study is carried out. However, if these experienced teachers, who also have been involved in the development of national tests, cannot reliably assess students' oral performances under the premises of the described assessment situation, then it does not seem likely that the group of all Swedish mathematics teachers, with very different experiences, would be able to use these scoring rubrics in a national test situation to score their students' performances with high interrater reliability. However, since the judgements of the students' subject matter knowledge displayed even lower interrater reliability the two oral communication aspects do not in themselves seem to have been the decisive factors for the somewhat low interrater reliability.

The study indicates that the assessment model under investigation does not provide sufficient guiding for how to interpret the constructs supposed to be rated and how to distinguish the levels of the constructs described by the scoring rubrics. Given that no descriptions of the constructs and only short descriptions of the rating categories were provided this should come as no surprise. However, this conclusion is no less important since there are other assessment models in different contexts that still use this level of description of constructs and rating categories.

The national system in which the assessment model functions may make it difficult to develop test material that sufficiently well supports the raters in their judgements of the students' performances. The descriptions of the different scoring levels are based on an interpretation of the national grading criteria, and these very generally written criteria may be difficult to transform into distinct descriptions that can be used to reliably differentiate between performances. In particular, for the highest test grade, Passed with special distinction, there were no national grading criteria at all and therefore the national tests did not provide criteria for this performance level either. However, it would be useful to conduct a study of teachers' interpretation of student performances in relation to the scoring rubrics. Such a study could investigate in which ways the teachers interpret them differently and provide information about the characteristics of the interpretation difficulties. The test material from 2003 and onwards include further developed scoring rubrics that could be used to investigate the difference in interrater reliability due to different characteristics of scoring rubrics. The new test material also employ new technologies that have been around for some years now that provide judges with oral exemplars in addition to written transcripts of oral performances. This new test material, which may provide sufficient information for teachers to attaining high agreement in their ratings, may be used to study how different characteristics of the descriptions of constructs and rating categories influence interrater reliability.

A possible factor that could negatively affect the interrater reliability is fatigue of the raters. Indeed, for the subaspect Line of thought the interrater reliability was significantly lower for the last three student performances than for the first three student performances, and fatigue may have played a role in this difference. Although this difference may have other causes, such as these performances may have been more difficult to rate in relation to this subaspect, it may be appropriate to investigate the approximate number of students that raters can assess in a row with preserved concentration (even though this number will vary considerably between raters). However, since the interrater reliability for both the first three and the last three student performances is not very high for any of the subaspects the somewhat low interrater reliability displayed in the study does not seem to mainly come from fatigue of the raters.

But, some other factors remain likely to have been important for the low interrater reliability. The teachers' limited confidence in their ratings may indicate that they had to little information for making confident ratings, or that they were uncertain of how to use the available information, that is how to interpret the scoring rubrics in relation to the students' performances. Insufficient information may be due to the very short time duration for the students' oral performances. In a pursue of making it feasible for teachers to assess all of their students' oral mathematical communication abilities (instead of a sample, which is done in for example Denmark and Norway) an attempt was made by the national test developers to minimise the time the teachers need to employ for the assessment. It may be that 5 minutes is just to little time for reliable scoring. In this time frame a few separate pieces of information may be decisive for the rating and if this information is difficult to interpret or are of a kind that corresponds to the border between two scoring levels then coincidences may strongly influence the scoring decisions and produce low interrater reliability. If this is so, then more time for the performances would be required for high interrater reliability, which however makes it less viable to carry out the assessment for the teachers in the schools. Thus, it would be useful to conduct a study investigating the influence of this time factor on the interrater reliability. The current national assessment procedure described earlier (in practice since 2003) allots 10 minutes for students' oral performances and may be used for investigating the time factor's influence on the interrater reliability.

Another factor that may have influenced the interrater reliability was the use of different tasks to solve and orally present the solution to. If many of the oral performances are based on the same task the teachers may administrate the oral performances so that students that have worked on the same task are assessed on their mathematical communication abilities at the same session. Assessing oral performances based on the same task may make it easier to reliably distinguish differences in the performances. However, this may bring about some disadvantages as well. For example, it will put limitations on the possibilities to choose tasks of suitable degree of difficulty for each student. It may also make it more difficult to administer the assessment if keeping students from working together with their planning of the oral performance is of importance (which it may be when individual abilities are assessed).

Yet another factor that may influence interrater reliability that has been suggested in the literature (although with ambiguous results) is training of the raters. Although this could be done, and might have some affect, it would be a major and costly implementation to train all Swedish mathematics teachers in this respect, or to develop a system of a smaller number of censors that travels around the country participating in the assessment (as in Denmark and Norway). However, if it is not possible to increase the interrater reliability through the test material and test procedure discussed above, and given that the assessment of oral mathematical communication is considered important to assess (with high interrater reliability), it may be necessary to investigate the effect of rater training on the interrater reliability of this kind of assessment of oral communication abilities.

Thus, oral mathematical communication has been argued to be an important part of mathematics and mathematics education and it is emphasized in all of the Swedish mathematics syllabi for upper secondary level. Therefore, the inclusion of this aspect in the national tests indeed does improve the alignment between the national syllabi documents and the national test system. However, this study shows that the oral part of the national test analysed in this study suffers from interrater reliability problems, and therefore interpretation and use of the test scores from this part would have been problematic. But the design of the study also allowed for comparison with the interrater reliability of the judgements of the subject matter knowledge displayed in the oral performances and from this data it can be concluded that it was not the oral mathematical communication aspects in themselves that made it difficult to reliably rate the performances. In the pursuit of deeper insights about the conditions required for consistency in the marking the influence of the clarity of the descriptions of the different performance levels in the scoring rubrics and accompanying support material, the time duration for the students' assessed performances, the variability coming from students' presenting solutions to different tasks, and the training of raters seem to be worth further study. Both when it comes to the development of the oral part of the Swedish national test for course C and the assessment of oral mathematical communication in general, investigations of the influence of these factors on the interrater reliability may be fruitful ways to gain important insights that can be used for decisions when trade-offs have to be made between high interrater reliability in the assessment of oral mathematical communication abilities on the one hand and costs and resources on the other hand.

References

Barnes, M, Clarke, D. & Stephens, M. (2000). Assessment: the engine of systemic curricular reform? *Journal of Curriculum Studies*, 32 (5), 623–650.

Brennan, R.L. & Johnson, E.G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, (14) 4, 25–27.

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin*, 70, 213–220.

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

Danish Ministry of Education (2007). *Matematik Stx*. Retrieved June 12, 2008 from http://us.uvm.dk/gymnasie/fagene/matematik/ny_stx.htm?menuid=15l

Finnish National Board of Education. (2004). *National core curriculum for upper secondary schools 2003*. Vammala: Finnish National Board of Education.

Frick, T. & Semmel, M.I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research*, 48, 157–184.

Glass, G.V. & Hopkins, K. H. (1996). *Statistical methods in education and psychology*. Boston: Allyn and Bacon.

Harman, H.H. (1967). *Modern factor analysis*. University of Chicago press.

Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education*, (23) 4, 367–378.

Kane, M., Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18 (2), 5–17.

Linacre, J.M. (1994). *Many-facet Rasch measurement*. Chicago: MESA press.

Linacre, J.M. (2002). Judge ratings with forced agreement. *Rasch Measurement Transactions*, 16 (1), 857–858.

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher*, 20 (8), 5–21.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (Vol. 3, pp. 13–103). New York: American Council on Education.

NCTM (2000). *Principles and standards for school mathematics*. Reston, Va.: National Council of Teachers of Mathematics.

Niss, M., & Jensen, T. H. (2002). *Kompetencer og matematiklaering* [Competencies and mathematical learning] (No. 18). København: Undervisningsministeriets forlag.

Norwegian Directorate for Education and Training (2007). *Natural Science and Mathematics Studies*. Retrieved June 12, 2008 from http://www.udir.no/templates/udir/TM_Artikkel.aspx?id=3587

Nyström, P. (2004). Reliability of educational assessments: the case of classification accuracy. *Scandinavian Journal of Educational Research,* 48 (4), 427–440.

OECD (1999). *Measuring student knowledge and skills. A new framework for assessment.* Paris: OECD, Organisation for Economic Co-operation and Development.

Raymond, M.R. & Viswesvaran, C. (1991). *Least-squares models to correct for rater effects in performance assessment* (ACT research report series 91-8). (ERIC, No. ED344947)

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability theory: a primer.* Newbury Park, CA: Sage Publications.

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation,* 9 (4). Retrieved August 16, 2007 from http://PAREonline.net/getvn.asp?v=9&n=4

Svensson, E. (1998). Ordinal invariant measures for individual and group changes in ordered categorical data. *Statistics in Medicine,* 17 (24), 2923–2936.

Swedish National Agency of Education (2001). *Natural science programme: programme goal, structure and syllabuses*. Stockholm: National Agency of Education

Swedish Ministry of Education and Research (1993). *Proposition 1992/93:250. Ny läroplan och ett nytt betygssystem för gymnasieskolan, komvux, gymnasiesärskolan och särvux*. Stockholm: Fritzes.

Tierney, R. & Simon, M. (2004). What's still wrong with rubrics. Focusing on consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation,* 9 (2), Retrieved August 16, 2007 from http://pareonline.net/getvn.asp?v=9&n=2

Uebersax, J. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin,* 101 (1), 140–146.

Uebersax, J. (2007). *Statistical methods for rater agreement.* Retrieved April 24, 2007, from http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm

Webb, N.L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph no. 6). Madison, WI: National Institute for Science Education.

Wiggins, G. (1998). *Educative assessment: designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass Publishers.

Wiliam, D. (2007). Keeping learning on track: classroom assessment and the regulation of learning. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1051–1089). Charlotte, NC: Information Age Publishing.

Torulf Palm

Torulf Palm, PhD, is a member of *Umeå mathematics education research centre* (UMERC) at Umeå University, Sweden and works at the department of *Mathematics, technology and science education.* His special research interests are assessment, mathematical reasoning, mathematical modelling and the authenticity of word problems.

torulf.palm@math.umu.se

Sammanfattning

Matematisk kommunikation, muntlig och skriftlig, ses i allmänhet som en viktig aspekt av matematik och matematikutbildning. Detta medför att muntlig matematisk kommunikation också bör bedömas i olika sorters prov. I både utbildningssammanhang och inom andra områden är dock muntliga prov av ämneskunskap eller kommunikationsförmåga behäftade med reliabilitetsproblem vilket orsaker svårigheter med dess användning. När det gäller matematikutbildning är reliabilitetsstudier om muntliga prov dock relativt ovanliga, och detta gäller speciellt vid bedömning av matematisk kommunikationsförmåga. Denna studie analyserar interbedömarreliabiliteten för bedömningen av muntlig matematisk kommunikation i ett svenskt nationellt matematikprov för gymnasial nivå. Resultaten visar att provet var behäftat med reliabilitetsproblem. Det verkar dock inte vara kommunikationsaspekten i sig själv som gör att denna förmåga var svår att bedöma reliabelt utan otillräckligheter i den då använda bedömningsmodellen.