A peer reviewed, open-access electronic journal: ISSN 1531-7714

How Raters Differ: A Study of Structured Oral Mathematics Assessment

Samuel Sollerman, *Stockholm University*

Abstract: This study examines the nature and extent of interpretive variability in structured oral mathematics assessments. Using Swedish national test data from 74 students across three oral formats, six experienced teachers independently rated reasoning, communication, and method using shared rubrics. Multiple reliability indicators and Svensson's method were employed to distinguish systematic and unsystematic interpretive variation. Exact agreement was low across formats, with higher but still modest adjacent agreement. Relative Position effects were frequent, indicating systematic differences in rater thresholds. In contrast, the most dialogic format showed greater Relative Rank Variance, suggesting more random inconsistency. Raters reported high confidence even when statistical agreement was low, revealing a gap between perceived certainty and interpretive alignment. The analysis indicates that assessment structure and interactional demands shape both what students display and how raters apply criteria, making variability a feature of professional judgment rather than merely error. Implications include the use of calibrated exemplars, targeted calibration activities, and collaborative scoring practices to enhance reliability without sacrificing the diagnostic value of oral assessment in competency-based systems.

Keywords: Oral assessment; Inter-rater reliability; Rater judgment; Performance-based assessment; Structured assessment formats; Mathematics education

Introduction

Assessing student performance—particularly in tasks involving reasoning, problem-solving, or oral communication—often depends on subjective human judgment (Brookhart, 2013; Black & Wiliam, 1998). Even when scoring guides are standardized, raters may interpret student responses differently, leading to variability in scoring (Brookhart, 2013; Palm, 2008a). These discrepancies pose challenges for fairness, validity, and trust in assessment systems (Nitko & Brookhart, 2011; Rudner, 1992). At the same time, international large-scale assessments such as PISA and TIMSS often report high inter-rater reliability in mathematics scoring (von Davier et al., 2024; OECD, 2024), largely due to the subject's focus on objectivity

and the dominance of standardized written formats (OECD, 2018; OECD, 2023; Mullis et al., 2020). Yet research has questioned how well these constructs align with the goals of national curricula and pedagogical practices (Sollerman, 2019). Moreover, the growing policy emphasis on authentic and performance-based assessments in national and local contexts raises new questions about how reliably complex student reasoning can be judged when responses are dialogic, spontaneous, and context-dependent (OECD, 2018; Phung & Michell, 2022; Cole, 2023).

Compared to written tests, oral assessments are more vulnerable to interpretive variability (Palm, 2008b; Joughin, 1998; Lind Pantzare, 2015). They are valued for capturing students' real-time reasoning and mathematical communication, but they rely on interactional cues and spontaneous dialogue, which make consistent scoring more difficult. Structured rubrics and training can mitigate such issues, but research indicates that raters may still diverge even when using the same criteria (Rudner, 1992; Nitko & Brookhart, 2011). Recent methodological developments emphasize the need for systematic approaches to documenting and improving inter-rater reliability across diverse assessment contexts (Cole, 2023).

As authentic and performance-based assessment practices expand and new technologies are explored to support large-scale scoring, the question of human rater reliability becomes increasingly urgent (Hwang et al., 2020; Chen et al., 2020). Before any large-scale or technology-supported scoring systems can be trusted, human raters themselves must demonstrate consistency and fairness in applying the underlying rubrics. Such variation is of particular interest because understanding its nature and extent is critical for developing valid and equitable assessment systems.

Building on these challenges, the present study investigates the nature and sources of rater disagreement in structured oral mathematics assessments. The study draws on data from the Swedish national mathematics course tests (Mathematics 1a–c, 3b–c, and adult education 1b) administered in autumn 2014, which have since been publicly released. In Sweden, the oral subtests of the national mathematics course tests were designed to assess multiple mathematical competencies, with particular emphasis on reasoning and communication abilities that are increasingly emphasized in national curricula. In the present study, these tests provide an authentic yet structured context for exploring rater agreement.

Research Objectives and Questions

The overarching aim of this study is to investigate the nature and extent of rater divergence when assessing student performance in structured oral mathematics assessments. The study aims to identify patterns and sources of disagreement among raters, rather than to establish causal explanations. This work also aims to inform more equitable assessment practices—both for individual learners and within large-scale evaluation systems—by clarifying how interpretive variability emerges and can be addressed in practice. Specifically, it aims to quantify the degree of inter-rater agreement across three distinct oral assessment formats, differentiate between systematic and unsystematic disagreement using Svensson's method, and examine how rater confidence relates to actual agreement levels. In this article, the term *interpretive variability* is used to describe these systematic and unsystematic differences in how raters apply shared criteria when judging student performance.

By addressing these objectives, the study contributes to a deeper understanding of how professional judgment operates in authentic assessment contexts and informs the design of scoring systems and rater training that support greater reliability, transparency, and fairness.

The following discussion situates these objectives within the broader theoretical and policy context of oral mathematics assessment.

Background

As discussed by Joughin (1998), oral assessment formats are used across subjects and contexts to evaluate students' ability to communicate knowledge in real time, reason spontaneously, and engage in dialogic exchange. These formats offer pedagogical depth but present challenges for consistent scoring, especially in large-scale or high-stakes settings (Joughin, 1998). In mathematics, oral assessment can reveal forms of reasoning and adaptability that written formats often overlook (Palm, 2008b). However, the interpretive demands of these formats raise concerns about reliability, particularly when raters must judge spontaneous and often nuanced responses that unfold through interactive dialogue rather than fixed written products. Recent research continues to emphasize that such interpretive demands and resulting variability pose one of the most persistent challenges for reliable assessment in mathematics (Roos & Bagger, 2024).

This challenge is compounded by the fact that the existence of scoring rubrics alone does not guarantee reliability. As Nitko and Brookhart (2011) point out, rater agreement depends on shared interpretive understanding and calibration, not just on access to criteria. Rudner (1992) highlights this further by identifying common sources of scoring error, such as leniency, severity, and halo effects, and recommends multiple strategies for reducing these inconsistencies, including rater training and statistical feedback. Research has also shown that raters may express high confidence in their judgments even when the level of agreement is objectively low (Palm, 2008b), suggesting that interpretive variability may go unnoticed. Such undetected divergence is particularly concerning in oral settings, where performance and assessment co-evolve in real time.

In parallel, research on students' mathematical reasoning has highlighted how beliefs, expectations, and affective factors shape the ways students present their thinking during assessments. Sumpter (2013) demonstrated that upper-secondary students frequently draw on notions of safety, motivation, and perceived competence when deciding how to reason and explain. As Sumpter (2013) demonstrated, students' reasoning choices are shaped by their sense of safety, motivation, and perceived competence—factors that may also influence how raters interpret their performance. In oral assessments, these socio-cognitive influences interact with the immediacy of dialogue and feedback, complicating the task of interpreting student performance and increasing the likelihood of rater divergence. Given these challenges, it is important to better understand how features of oral assessments may influence how raters interpret and apply the criteria. Jönsson and Balan (2017) offer insight into how analytic and holistic rubric structures may differently affect inter-rater reliability, suggesting that the format and cognitive demands of the task interact with the scoring model to shape rater judgment. Recent findings similarly show that differences in teachers' assessment accuracy and interpretive framing can have measurable implications for student outcomes and reliability (Kolovou et al., 2024). This interaction points to the need for research that examines how oral assessment design and rater interpretation jointly contribute to reliability outcomes.

To better understand the nature of oral assessments and their implications for inter-rater reliability, Joughin's (1998) framework offers a useful analytical lens. He describes oral assessment along five dimensions: primary content focus (e.g., knowledge, reasoning, or application), interaction pattern (e.g., monologue vs. dialogue), authenticity, structure (scripted vs. open), and the degree of examiner judgment required. These dimensions help differentiate oral assessment types and illuminate how specific formats may foster more subjective or variable interpretations by raters. In other words, the structure and interactional mode of the assessment shape not only what students demonstrate but also how reliably teachers can evaluate it.

Building on this conceptualization of assessment formats, it becomes important to apply methods that can reveal not just the presence but also the nature of disagreement between raters. In light of the format-related variation outlined above, ensuring reliability in assessment is essential for maintaining fairness and trust in evaluation systems. Recent policy frameworks, such as the OECD's (2023) focus on reasoning and

communication in PISA 2022, further underline the need for assessment formats that are both valid and reliable when evaluating complex competencies. While quantitative methods, such as Cohen's kappa and percent agreement, provide overall indicators of agreement, they do not distinguish between systematic bias (e.g., a rater consistently more lenient) and random noise (inconsistency). Svensson's method adds granularity by enabling the identification of both types of disagreement (Svensson, 2012). Applying such an approach allows for a more diagnostic understanding of reliability—one that clarifies whether disagreement reflects structural, interpretive, or purely random variation.

While previous research has identified challenges of reliability in oral assessment (e.g., Palm, 2008b; Joughin, 1998), less is known about how structural differences between oral formats contribute to distinct patterns of rater disagreement, particularly in mathematics education

In this study, Svensson's method is used to explore not only the extent but also the underlying patterns of rater disagreement in structured oral assessments. By comparing three distinct oral assessment formats, we examine how structural features—such as task design, interaction pattern, and scoring criteria—interact with rater interpretation. Through this lens, the study seeks to connect theoretical insights on assessment format (Joughin, 1998) with empirical measures of disagreement (Svensson, 2012), providing a bridge between conceptual and methodological perspectives. These considerations provide the conceptual foundation for the present study, which examines rater interpretation across three oral assessment formats.

Study Context

Understanding these dynamics requires attention to the policy context in which such assessment practices emerged (e.g. Imsen et al., 2016). Educational reforms across the Nordic region have introduced new accountability mechanisms alongside efforts to promote competencies such as reasoning, communication, and problem-solving (Niss & Højgaard, 2011; OECD, 2019; Skolverket, 2011; Imsen et. al, 2016). In recent years, these competency-oriented reforms have been further reinforced through policy frameworks that emphasize mathematical reasoning, problem-solving, and communication as key dimensions of proficiency (OECD, 2023; Skolverket, 2022). In the Nordic context, mathematics education research highlights how teachers negotiate policy demands, professional judgment, and equity concerns in their assessment practices (Roos & Bagger, 2024). Such reforms, reflecting deeper shifts in education governance, have created tensions between equity-based traditions and performance-based demands. Oral assessment formats can be seen as one response to these dual pressures.

Sweden offers a particularly relevant case within the Nordic context, as it implemented structured oral components in large-scale national mathematics assessments (Kjellström & Pettersson, 2005). Between 2011 and 2018, Sweden included structured oral parts in national assessments for upper-secondary mathematics. These assessments were centrally designed but scored locally by teachers using national rubrics. They were introduced in response to curricular demands emphasizing that teachers must make well-grounded evaluations of students' knowledge and that students should be able to communicate mathematical thinking orally, in writing, and in action (Skolverket [Swedish National Agency for Education], 2011). Since then, curriculum updates have continued to stress students' ability to reason and communicate mathematically as integral to both learning and assessment (Skolverket, 2022). Oral assessment was expected to enhance validity by capturing forms of mathematical competence, such as reasoning and adaptability, that written formats often miss (Palm, 2008b). Despite structured tasks and scoring criteria, research indicates considerable variability in how raters interpret and apply the rubrics (Palm, 2008b). This persistent interpretive variability makes the Swedish case particularly informative for understanding the reliability challenges that arise when competency-based assessment is implemented in practice.

To investigate these patterns in practice, we designed a study comparing three different oral assessment formats, drawing on both Joughin's dimensions and Svensson's analytical approach. By situating the analysis within this policy and curricular context, the study connects local assessment practices to broader international efforts to balance validity, reliability, and equity in mathematics education.

Method

The study was designed to compare three structurally distinct oral assessment formats in upper-secondary mathematics, each differing in interaction type, task design, and expected cognitive demands. This comparative design was chosen to capture how variations in structure and interaction, as described by Joughin's (1998) framework, may influence rater interpretation and agreement.

Assessment Formats

We included oral assessment tasks from three parts of Sweden's upper-secondary mathematics education: two tasks associated with the first mathematics course and one with the third course. One version of the first course was delivered in the regular upper-secondary school, while the other was part of municipal adult education. Although both addressed the same curriculum content, they differed in format due to contextual and logistical factors: the regular upper-secondary school version typically involved small-group interaction, while the adult education version required paired or individual assessment. The oral part, linked to the third mathematics course, was designed for more advanced students and focused on student-led presentations of extended problem-solving.

To reflect the structural and interactional differences among these formats—and drawing on Joughin's (1998) typology—we refer to them in this article as the Dialogic Group Format, the Scripted Individual Format, and the Dialogic Paired Format (see Table 1). As Joughin (1998) and Palm (2008a) suggest, more dialogic or less structured formats may place greater interpretive demands on raters, increasing the likelihood of disagreement.

Table 1. Overview of oral assessment formats in upper-secondary mathematics

Format Label	Source Course	Interaction Type	Structure	Judgment Demand	Example Focus
Dialogic Group Format	First course	Small-group, dialogic	Semi-structured	High	Student discussion guided by teacher
Dialogic Paired Format	First course	Paired, dialogic	Lightly structured	High	Evaluating/reflecting on solutions guided by the teacher
Scripted Individual Format	Third course	Monologic	Scripted	Moderate	Individual problem presentation

The three formats differed not only in structure and interaction but also in how they shaped the scoring context, as seen in Table 1. The Scripted Individual Format emphasized monologic student presentations, which offered greater control over content but limited opportunities for probing or clarification. In contrast, the Dialogic Group Format encouraged peer interaction in teacher-led discussions, introducing more variability but also richer data on reasoning. The Dialogic Paired Format combined dialogic interaction with reflective tasks, often requiring students to evaluate or critique existing solutions guided by the teacher.

While all formats assessed reasoning, communication, and procedural competence, the differences in structure and interaction likely contributed to varying interpretive demands on raters.

Participants and Rating Procedure

The study included 74 student performances: 26 in the Dialogic Group Format, 24 in the Scripted Individual Format, and 24 in the Dialogic Paired Format. Students were selected by their regular teachers, with the aim of capturing a broad range of performance levels and student backgrounds. The purpose of this sampling was not to achieve statistical representativeness, but to reflect authentic classroom variation across different upper-secondary contexts. Each student's performance was rated independently by six trained mathematics teachers. This number was chosen to balance feasibility and statistical power, consistent with common practices in inter-rater reliability research (Nitko & Brookhart, 2011; Bresciani et al., 2009). This design aligns with prior work on large-scale assessments where teachers score without external controls, raising questions about reliability (Lind Pantzare, 2015). The same six raters participated in the evaluation of all three oral assessment formats, ensuring comparability across contexts. The raters were qualified and experienced teachers with between 10 and 40 years of teaching experience and varied in age from just over 30 to over 60. Half were women representing municipal and independent schools and urban and rural areas. Their professional backgrounds included experience in both vocational and academic programs.

Ratings were conducted in connection with the actual test administration, as students' regular teachers led the sessions to replicate authentic classroom conditions. The rubrics, designed to guide consistent interpretation and support scoring reliability, were analytic in structure. In the Dialogic Group and Paired Formats, raters scored three dimensions—Method, Reasoning, and Communication—each on an ordinal three-level scale (E–C–A) reflecting increasing proficiency. The Scripted Individual Format included three subcriteria within a single communication domain (completeness and structure, explanations, and mathematical terminology). Across all formats, descriptors emphasized procedural accuracy in Method, logical coherence in Reasoning, and precision of mathematical language in Communication. Scores for each criterion were summed to yield a total performance score, ranging from 7 to 11 points depending on format. These structural differences in rubric design help explain the observed variation in rater agreement across formats.

Raters received written scoring instructions, but no calibration session or joint discussion of rubric interpretation occurred before the assessment. This was a deliberate design choice intended to maintain ecological validity by reflecting typical school practice, where oral assessments are usually conducted individually by the teaching teacher and/or a colleague without centralized training (Nordberg, 2018). While this may reduce inter-rater agreement, it offers valuable insight into how reliability challenges manifest under authentic assessment conditions. While this enhances ecological validity, it also introduces interpretive variability—a design feature that constitutes both a strength and a limitation of this study. The raters in this study observed and scored performances in real-time alongside the students' regular teachers. Raters were aware of student identities, a feature that can introduce scoring variation since prior perceptions of student ability may influence judgments (Meier et al., 2006). In addition to assigning scores, raters also indicated how confident they were in each judgment by responding to the prompt: *"How certain are you that the student should or should not receive the point?"* Confidence was recorded on a four-point Likert scale (Very uncertain, Somewhat uncertain, Somewhat certain, Very certain). This measure provided an additional indicator of raters' perceived certainty in their scoring decisions.

The study followed institutional ethical guidelines for the use of anonymized educational data. All participants provided informed consent, and student identities were not recorded in the analytical phase.

Analytic Strategy

The six raters formed 15 unique rater pairs for analysis, providing a robust basis for examining inter-rater agreement. To evaluate reliability, we drew on Stemler's (2004) three-part framework, which distinguishes between consensus, consistency, and measurement estimates. Consensus estimates assess the degree to which raters assign the same score. Consistency estimates reflect whether raters rank performances in a similar order. Measurement estimates evaluate the reliability of aggregated scores across raters. Svensson's method was selected because it provides diagnostic insight into the nature of disagreement—distinguishing between systematic rater tendencies and random variation—thereby aligning with the study's focus on how and where raters diverge in their judgments.

For consensus estimates, we calculated percent agreement and adjacent agreement. Percent agreement measures the proportion of exact score matches, while adjacent agreement includes scores within one or two points of each other. In educational contexts, a 70 percent agreement rate is often considered acceptable (Stemler & Tsai, 2008). We also applied Cohen's kappa to adjust for agreement expected by chance. Based on the thresholds from Landis and Koch (1977), values below 0.20 indicate slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, and values above 0.60 indicate substantial agreement. While these thresholds are commonly used, Stemler and Tsai (2008) note that in educational settings, a kappa of 0.50 or higher is often acceptable, especially when raters are evaluating complex student performances.

To estimate consistency, we used Spearman's rank correlation to evaluate whether raters ranked performances in a similar order. Correlations closer to 1.0 indicate stronger agreement. A value of 0.70 or higher is generally accepted as a benchmark for sufficient agreement in educational research (Multon, 2010).

In addition to these summary measures, we applied Svensson's method (Svensson, 2012), which is especially suited for ordinal data and qualitative judgments. Unlike global statistics like kappa, Svensson's method identifies whether disagreement stems from systematic rater patterns or from random inconsistency. The method includes three components: Relative Position (RP), which indicates systematic leniency or severity; Relative Concentration (RC), which reflects how differently raters spread their scores; and Relative Rank Variance (RV), which captures unsystematic variation. Significance for RP, RC, and RV was assessed using 95 percent confidence intervals.

Results

To understand both the extent and the nature of rater disagreement, the results are presented using multiple reliability measures. Consensus and consistency indicators (percent agreement, kappa, and Spearman) demonstrate that raters differ. At the same time, Svensson's method offers insight into how they differ, by distinguishing systematic tendencies from random variation in scoring patterns. To assess the extent of rater disagreement across formats, we first examined descriptive reliability measures, including percent and adjacent agreement, as well as score variation.

Percent agreement was low, ranging from 17% in the Dialogic Paired Format to 27% in the Scripted Individual Format. Adjacent agreement, which considers ratings within two points of each other as acceptable, was higher but still revealed significant inconsistency. To capture more nuanced levels of consensus, two adjacent-agreement thresholds were calculated: Δ_1 (scores differing by at most one point) and Δ_2 (scores differing by at most two points). Table 2 summarizes these indicators together with chance-corrected ($\varkappa$) and rank-order (ϱ) estimates.

Table 2. Descriptive reliability indicators across assessment formats

Format Label	Percent Agreement	Adjacent Agreement Δ_1	Adjacent Agreement Δ_2	Cohen's $\varkappa$	Spearman's ϱ
Dialogic Group Format	25 %	63 %	84 %	0.16	0.81
Dialogic Paired Format	17 %	55 %	78 %	0.06	0.62
Scripted Individual Format	27 %	64 %	87 %	0.07	0.40

Note. Δ_1 = agreement within one point; Δ_2 = agreement within two points. $\varkappa$ = Cohen's kappa (chance-corrected agreement); ϱ = Spearman rank correlation (rank consistency).

As shown in Table 2, both adjacent-agreement measures were substantially higher than exact agreement, indicating that most discrepancies involved small rather than large score differences. Nevertheless, even with this tolerance, agreement levels remained modest, particularly for the Dialogic Paired Format. Cohen's $\varkappa$ values (.06–.16) confirm that overlap was only slightly above chance, placing all formats within the *slight agreement* range according to Landis and Koch (1977) and below the 0.50 threshold often considered acceptable for complex educational tasks (Stemler & Tsai, 2008). This indicates a consistent problem with rater alignment even after adjusting for chance agreement, reflecting deeper interpretive divergence in scoring.

Spearman's ϱ values (.40–.81) suggest that raters shared a general sense of rank order but differed in the absolute scores they assigned. Taken together, these results show that broad consensus on overall performance was accompanied by instability in finer-grained distinctions—a pattern that points to interpretive ambiguity in how rubric criteria were applied. Such inconsistency implies that even when raters share an overall sense of proficiency, the scoring language may leave room for divergent judgments about what constitutes sufficient evidence within each category.

The maximum possible score differed across formats (11 points in the Dialogic Group Format, 7 in the Scripted Individual, and 9 in the Dialogic Paired). Average variation within rater pairs amounted to 13 % of the total score in the Dialogic Group Format and 18 % in both the Scripted Individual and Dialogic Paired Formats. These figures further illustrate the instability of fine-grained scoring even under structured assessment conditions.

To further examine the nature and extent of rater divergence, Svensson's (2012) method was applied to determine whether disagreement reflected consistent bias patterns or unsystematic scatter. It distinguishes between systematic disagreement (e.g., consistent leniency or severity) and unsystematic variation (random disagreement). This approach provides a diagnostic complement to the descriptive reliability indicators by separating three components of rater disagreement. High Relative Position (RP) values indicate that one rater consistently scores higher or lower than another—reflecting bias in leniency or severity—while Relative Rank Variance (RV) captures unpredictable divergence between raters who otherwise follow similar tendencies. Relative Concentration (RC) reflects differences in score spread, showing whether a rater tends to cluster scores more narrowly or broadly. This diagnostic structure allows the analysis not only to quantify disagreement but to reveal its character—systematic or random—across assessment formats.

In total, 11 of the 15 rater pairs in both the Dialogic Group and Scripted Individual Formats, and 12 of 15 pairs in the Dialogic Paired Format, showed significant Relative Position (RP) values. The present of Relative Concentration (RC) was low in all formats. Significant Relative Rank Variance (RV) values were

found in 9 pairs for the Dialogic Group and Dialogic Paired Formats and in 6 pairs for the Scripted Individual Format (see Table 2).

The pattern of results in Table 3 indicates that systematic differences (RP) were more frequent than random ones (RV), suggesting that bias in leniency or severity was the dominant source of rater divergence. However, the Dialogic Paired Format also shows high RV values, implying that open, dialogic structures generate greater random inconsistency.

Table 3. Full Results of Svensson's Method for All Rater Pairs

Rater pair	Dialogic Group Format			Scripted Individual Format			Dialogic Paired Format		
	RP	RC	RV	RP	RC	RV	RP	RC	RV
R1 vs R2	-0.354*	-0.280	0.096*	0.349*	-0.236	0.228*	-0.434*	0.070	0.126
R1 vs R3	-0.478*	-0.373	0.090	0.297*	-0.303	0.242*	-0.530*	0.019	0.090
R1 vs R4	-0.423*	-0.281	0.139*	0.464*	-0.310	0.032	-0.531*	-0.579	0.305*
R1 vs R5	-0.444*	-0.906	0.082	-0.078	-0.055	0.048	-0.698*	-0.816	0.141
R1 vs R6	-0.411*	0.015	0.131	0.677*	-0.357	0.156*	-0.705*	-0.993	0.234*
R2 vs R3	-0.408*	-0.278	0.053	-0.017	-0.088	0.332*	-0.175	0.146	0.036
R2 vs R4	-0.327*	-0.046	0.083*	0.137	-0.047	0.431*	-0.250*	-0.294	0.319*
R2 vs R5	-0.370*	-0.441	0.031*	-0.394*	0.150	0.286	-0.424*	-0.195	0.086
R2 vs R6	-0.308*	0.121	0.051	0.396*	-0.053	0.294	-0.444*	-0.311	0.140
R3 vs R4	-0.101	0.249	0.091*	0.142	0.054	0.281	-0.226	-0.571*	0.260*
R3 vs R5	-0.145	0.091	0.059*	-0.344*	0.256	0.292	-0.474*	-0.851*	0.083
R3 vs R6	-0.077	0.313*	0.122*	0.378*	0.072	0.338*	-0.440*	-0.700*	0.088
R4 vs R5	-0.306*	-0.353	0.091*	-0.498*	0.260	0.055	-0.293*	0.079	0.371*
R4 vs R6	-0.219*	0.083	0.161*	0.262*	0.010	0.054	-0.301*	-0.037	0.194*
R5 vs R6	-0.197	0.328*	0.113	0.700*	-0.446	0.095	-0.161	-0.171	0.228

Note. R1 to R6 denote the six different raters. Statistically significant values ($p < .05$) are marked with an asterisk

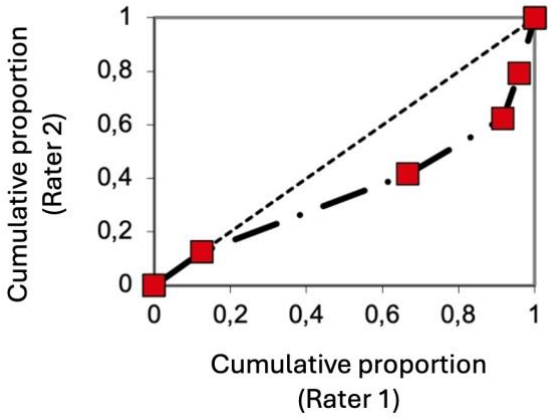
As can be seen in Table 3, several rater pairs showed notably high values on one or more Svensson indicators. For example, RP values reached up to 0.71, RC up to 0.99, and RV up to 0.43, underscoring the presence of both systematic and unsystematic disagreement. Notably, Relative Position (RP) values highlighted systematic divergence in several rater pairs, especially in Scripted Individual Format, while Dialogic Paired Format displayed both high RP and Relative Rank Variance (RV), indicating a mix of bias and inconsistency. Together, these findings show that disagreement was not random but structured differently across formats, with structured formats producing predictable leniency or severity and dialogic formats yielding greater random inconsistency.

Svensson's method also includes visual tools such as Relative Operating Characteristic (ROC) curves, which display the cumulative proportions of student scores assigned by two raters. These are visualisations based on the cumulative proportions of student scores between two raters. A diagonal line indicates perfect agreement, while deviation from this line reveals systematic differences in how scores are distributed. These curves provide an intuitive picture of bias, complementing the RP metric. In this study, ROC curves served as intuitive diagnostic complements to the numerical indicators: the degree and direction of curve deviation illustrated whether disagreement stemmed primarily from consistent bias (RP) or random scatter (RV).

Svensson's method also includes visual tools such as Relative Operating Characteristic (ROC) curves, which display the cumulative proportions of student scores assigned by two raters. A diagonal line represents perfect agreement, while deviation from this line illustrates systematic shifts in scoring distributions (RP). Irregularities in the curve's shape can further hint at random variation (RV). These curves therefore provide an intuitive visual complement to the numerical indicators, helping to distinguish between patterned and random disagreement in scoring. In this study, ROC curves served as diagnostic aids to interpret whether observed divergence primarily reflected consistent interpretive differences or random inconsistency.

An example from the Scripted Individual Format is shown in Figure 1, illustrating how systematic and random disagreement appear in the ROC visualization.

Figure 1. Relative Operating Characteristic (ROC) Curve Comparing Two Raters' Score Distributions



Note. The x-axis shows the cumulative proportion of student scores assigned by Rater 1, and the y-axis shows the cumulative proportion assigned by Rater 2. The diagonal represents perfect agreement. Deviations from this line illustrate systematic differences in score distributions between raters.

As illustrated in Figure 1, the curve deviates below the diagonal, indicating that Rater 2 consistently awarded higher scores, reflecting a more generous scoring tendency (RP = 0.35) and a systematic difference in how the raters interpreted performance levels. The irregular shape of the curve near the extremes also indicates unsystematic variation (RV = 0.228). This pattern illustrates how ROC curves can reveal both structured bias and random variation in rater disagreement, complementing the numerical indicators. This visualization also highlights that while some divergence may be reduced through clearer rubrics or training, other sources of variability appear less controllable and may reflect external factors or task design limitations.

Beyond statistical indicators, further patterns emerged when we examined scoring confidence and specific assessment criteria. However, further patterns emerged when examining disagreement across specific scoring dimensions. Disagreement was most pronounced on qualitative criteria such as Reasoning and Communication. At the course level, the Dialogic Paired Format showed the lowest percent agreement and the highest Relative Rank Variance. At the same time, the Scripted Individual Format displayed the most prominent systematic bias. These patterns help clarify where disagreement was most evident, across both

scoring criteria and course contexts. These format-specific trends suggest that reliability challenges are closely linked to the interpretive demands of each assessment design. These aspects depend heavily on the subjective interpretation of student responses, even when rubrics are available. More procedural components, such as Method, showed relatively higher consistency among raters.

These format- and criterion-level findings provide important context for understanding rater behavior more broadly. Raters also completed a four-point scale indicating how confident they felt in their assessments. Although measurable disagreement, many reported high certainty in their ratings. On average, raters reported feeling 'quite certain' or 'very certain' about 92% of their judgments in the Dialogic Group Format, 85% in the Scripted Individual Format, and 90% in the Dialogic Paired Format. These figures indicate that raters often reported high confidence in their scoring, even when the statistical agreement was low. This gap between perceived certainty and actual statistical agreement highlights a metacognitive dimension of reliability: raters may not be aware of their own interpretive divergence.

Taken together, these results address the research question by showing how and where rater disagreement occurs. Systematic leniency or severity dominated in structured, monologic tasks, whereas dialogic and less scripted formats introduced greater random divergence linked to interpretive demands. These findings provide a basis for the subsequent discussion; implications for assessment design and professional judgment are considered in relation to fairness and trust in competency-based evaluation systems.

Discussion

These results highlight that disagreement among raters is not only widespread but also patterned across formats, scoring dimensions, and raters' own perceptions of certainty. This study confirms that oral assessment in mathematics—an important tool for accessing students' higher-order competencies—poses substantial challenges for inter-rater reliability (Palm, 2008b; Joughin, 1998). Svensson's method provides a powerful lens for detecting both systematic and unsystematic forms of interpretive variability that might be overlooked in more general reliability metrics. The findings indicate that disagreement was not evenly distributed but patterned across both formats and scoring dimensions. Qualitative criteria like reasoning and communication were especially prone to variation, consistent with prior findings (Pettersen & Nortvedt, 2018). Some formats encouraged more open-ended responses, which increased interpretive demands. For instance, the Dialogic Paired Format showed both systematic and unsystematic interpretive variability, while the Scripted Individual Format revealed variation in rater thresholds. These results therefore clarify how and where rater divergence occurs—across assessment formats that vary in structure and interaction rather than in the content being assessed. These structured patterns of interpretive variability suggest that format design influences both student responses (Sumpter, 2013) and rater interpretation (Joughin, 1998; Pettersen & Braeken, 2019).

While the Dialogic Group Format posed the greatest interpretive demands for raters, it also represented the assessment type most aligned with current curricular aims emphasizing reasoning and communication. In dialogic settings, students are required to articulate, challenge, and refine ideas in interaction with others—processes central to mathematical proficiency but difficult to capture in scripted formats. This creates a fundamental tension between validity and reliability: dialogic formats better reflect the competencies that curricula seek to promote, yet their openness makes consistent scoring more challenging.

Interpretive complexity is especially high in dialogic formats that require spontaneous interaction and flexible reasoning. These conditions make it more challenging for raters to apply rubrics consistently and may increase their reliance on assumptions or prior expectations. Additionally, the absence of calibration may have exacerbated variation in rater judgment. As Meier et al. (2006) note, prior familiarity with students

and differences in professional experience can lead to scoring bias. Even well-structured tasks are vulnerable to interpretive variability without shared training or norms. These structural features likely contribute to both systematic and unsystematic interpretive variation observed across contexts. Rather than suggesting causal mechanisms, these patterns point to specific sources of variation in rater interpretation, such as interactional openness, task design, and the granularity of rubric descriptors.

Future research should investigate how different oral formats—particularly dialogic ones—could be adapted or supported to reduce interpretive divergence while preserving their diagnostic value. In doing so, studies may further illuminate the relationship between assessment structure and professional judgment identified here.

These patterns raise important concerns for educational policy and practice (Black & Wiliam, 1998). Scoring disagreements in oral assessment are not just technical issues but have implications for fairness and equity (Rudner, 1992). If oral assessments are to support competency-based education effectively, they must be designed to ensure both validity and reliability. The mismatch between raters' high confidence and their actual levels of agreement underscores the need for calibration tools and shared frameworks. Such calibration does not necessarily require standardization but rather collective interpretation of scoring criteria through exemplars, discussion, and feedback—approaches shown to improve rating accuracy (Kolovou et al., 2024).

These supports are feasible through practices, such as annotated rubrics and digital feedback systems. Ensuring reliability in oral assessment will require both technical supports and professional learning structures. If oral assessments are to play a greater role in evaluating student achievement, particularly in areas such as reasoning and communication, they must be designed and implemented in ways that are not only valid but also reliable. The observed gap between perceived certainty and statistical agreement further emphasizes the importance of calibration and shared interpretive frameworks. This need aligns with professional testing standards that highlight rater training, rubric clarity, and ongoing validation as central components of reliable performance assessment (AERA, APA, & NCME, 2014). Moreover, the observed interpretive variability among raters reflects a broader professional challenge discussed in recent work on teacher judgment and assessment accuracy (Kolovou et al., 2024; Cole, 2023). Variability is not merely error but a manifestation of interpretive divergence—the professional negotiation of meaning that underpins the assessment of complex reasoning.

Improving reliability in oral assessments requires a combination of strategies, although this study did not directly test these interventions. Prior research has shown that even structured rubrics can lead to variability without shared calibration (Bresciani et al., 2009; Lind Pantzare, 2015). Practices such as annotated rubrics, rater training, and collaborative scoring could help reduce interpretive variation. For example, teacher reports suggest that co-assessment practices—ranging from shared observation to consensus discussions—are already being used with some success in Swedish mathematics education (Nordberg, 2018). These practices may offer models for broader application. The present findings provide empirical support for such initiatives by showing how variation emerges across different oral formats and how structured dialogue among raters could target both systematic and random forms of interpretive variability. Developing professional learning communities where teachers jointly interpret performance examples would help operationalize the kind of calibration emphasized in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014).

These steps could support a more reliable assessment without undermining the formative and diagnostic value of oral examinations. The high confidence expressed by raters—even in cases of statistical disagreement—highlights a challenge: evaluators may not be aware of their own interpretive variability. Confidence ratings in this study revealed a gap between subjective certainty and actual agreement, suggesting limited awareness of divergence. This underscores the importance of reflective assessment training

(Brookhart, 2013). As Jönsson and Balan (2017) found, scoring models and rubric design affect inter-rater agreement. Inter-rater reliability involves multiple dimensions—consensus, consistency, and measurement (Stemler, 2004; Stemler & Tsai, 2008)—and each may be influenced by how rubrics are interpreted. Svensson's method helps distinguish these layers and can support refinement of both tools and professional judgment. Future implementations could therefore use Svensson's indicators diagnostically, not only to measure reliability but to inform targeted rater feedback and validation processes.

These challenges are reinforced by findings from Pettersen and Braeken (2019), who argue that rubric reliability depends on alignment with actual task demands. Even well-designed rubrics may fail if raters do not share a clear interpretive understanding (Nitko & Brookhart, 2011). Systematic variation in interpretive thresholds—formerly described as leniency or severity—may partly explain the patterns seen in reasoning and communication criteria. The frequent occurrence of Relative Position differences in this study suggests that rater tendencies shaped scores in meaningful ways, even with structured criteria. This pattern underscores the need to view disagreement not only as error but as information that can guide refinement of rubrics and validation procedures. Viewed through this lens, divergence is not merely a function of individual bias but reflects variation in interpretive thresholds shaped by task structure, rubric design, and contextual expectations.

Meeting these challenges is both feasible and essential if oral assessments are to serve as fair, trustworthy indicators of student competence. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) emphasize that reliability and validity must be supported through iterative validation—where scoring criteria, rater behavior, and intended constructs are examined together. This perspective resonates with recent discussions of professional judgment as a balance between autonomy, fairness, and accountability in Nordic mathematics education (Roos & Bagger, 2024). As in other studies of teacher judgment accuracy (Kolovou et al., 2024; Cole, 2023), interpretive variability is shown here to be a condition of professional interpretation rather than its failure. At a system level, addressing interpretive variability is essential for maintaining fairness, transparency, and public trust in competency-based assessment reforms. The task, then, is not to eliminate variability but to manage it transparently through calibration, shared exemplars, and reflective practice. Such an approach would help reconcile validity and reliability in competency-based assessment, supporting both professional trust and system-level fairness.

Although the present design, with raters crossed with student performances, would in principle allow for a generalizability analysis, the study focused on identifying patterns of interpretive variability rather than estimating variance components. Future research could extend this work by applying generalizability theory to partition variance attributable to raters, students, and formats. In addition, future studies could explore how different oral formats—particularly dialogic ones—might be adapted or supported to reduce interpretive variability while preserving their diagnostic value.

Conclusion

This study examined patterns of interpretive variability across raters when scoring structured oral mathematics assessments. While oral formats provide valuable opportunities to assess reasoning and communication, they also introduce interpretive challenges that affect reliability. Using Svensson's method, the analysis identified systematic and unsystematic patterns of interpretive variability, often linked to scoring dimensions, assessment format, and rater expectations. These findings show that rubric clarity alone is insufficient for consistent scoring. Structured, monologic tasks tended to produce systematic differences in interpretive thresholds, whereas dialogic and less scripted formats generated greater random divergence. Such variability reflects the interpretive demands inherent in assessing open, interactive reasoning—precisely the competencies emphasized in contemporary curricula.

The implications extend beyond technical reliability. By demonstrating how interpretive variability arises, the study provides a diagnostic basis for improving calibration, rubric design, and validation procedures in line with established professional standards (AERA, APA, & NCME, 2014). Supporting teachers in reflective and collaborative calibration practices could strengthen both validity and trust in competency-based assessment systems. Ultimately, fair and trustworthy oral assessment depends not only on improved instruments but also on shared professional judgment and sustained structures for collaboration.

Received: 6/22/2025. **Accepted:** 1/2/2026. **Published:** 1/8/2026.

Citation: Sollerman, S. (2026). How raters differ: A study of structured oral mathematics assessment. *Practical Assessment, Research, & Evaluation*, 31(1)(2). Available online: https://doi.org/10.7275/pare.3236

Corresponding Author: Samuel Sollerman, Stockholm University.
Email: samuel.sollerman@su.se

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bresciani, M. J., Zelna, C. L., & Anderson, J. A. (2009). Examining design and inter-rater reliability of a rubric measuring research quality. *Practical Assessment, Research & Evaluation, 14*(1). https://doi.org/10.7275/1w3h-7k62

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Brookhart, S. M. (2013). How to create and use rubrics for formative assessment and grading. Alexandria, VA: ASCD.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access,* 8, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Cole, R. (2023). Inter-Rater Reliability Methods in Qualitative Case Study Research. *Sociological Methods & Research, 53*(4), 1944–1975. https://doi.org/10.1177/00491241231156971

Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers & Education: Artificial Intelligence, 1*, 100001. https://doi.org/10.1016/j.caeai.2020.100001

Imsen, G., Blossing, U., & Moos, L. (2016). Reshaping the Nordic education model in an era of efficiency: Changes in the comprehensive school project in Denmark, Norway, and Sweden since the millennium. *Scandinavian Journal of Educational Research*. https://doi.org/10.1080/00313831.2016.1172502

Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education, 23*(4), 367–378. https://doi.org/10.1080/0260293980230404

Jönsson, A., & Balan, P. (2017). Analytic or holistic: A study of agreement between different grading models. *Practical Assessment, Research & Evaluation, 22*(2). https://doi.org/10.7275/mg59-xq60

Kjellström, K. & Pettersson, A. (2005). The curriculum's view of knowledge transferred to national tests in mathematics in Sweden. *Zentralblatt für Didaktik der Mathematik* 37, 308–316 (2005). https://doi.org/10.1007/BF02655817

Kolovou, D., Veldhuis, M., van der Kleij, F. M., & Eggen, T. J. H. M. (2024). Does teacher judgment accuracy matter? How differences in teachers' assessment accuracy relate to student learning outcomes. *Teaching and Teacher Education, 138*, 104555. https://doi.org/10.1016/j.tate.2024.104555

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lind Pantzare, A. (2015). Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls? *Practical Assessment, Research & Evaluation, 20*(9). https://doi.org/10.7275/y2en-zm89

Meier, S. L., Rich, B. S., & Cady, J. (2006). Teachers' use of rubrics to score non-traditional tasks: Factors related to discrepancies in scoring. *Assessment in Education: Principles, Policy & Practice, 13*(1), 69–95. https://doi.org/10.1080/09695940600563512

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2020). *TIMSS 2019 international results in mathematics and science*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2019/international-results/

Multon, K. (2010). Interrater reliability. In N. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 627–629). Sage Publications. https://doi.org/10.4135/9781412961288.n194

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Pearson.

Niss, M., & Højgaard, T. (2011). *Competencies and Mathematical Learning: Ideas and Inspiration for the Development of Mathematics Teaching and Learning in Denmark*. Ministry of Education.

Nordberg, C. (2018). *Nationella provet i matematik i årskurs 9, 2018* [The national test in mathematics for grade 9, 2018]. PRIM-gruppen, Stockholms universitet.

Nordberg, C., Pettersson, A., & Sollerman, S. (2021). Bedömaröverensstämmelse vid bedömning av elevernas prestationer på de skriftliga delproven i 2017 års nationella prov i matematik för årskurs 9 [Inter-rater agreement in assessing students' performances on the written parts of the 2017 national mathematics test in Grade 9]. Stockholm University, Department of Mathematics and Science Education. (Stockholm University Reports in Mathematics Education No. 9). http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-189762

OECD. (2018). *The future of education and skills: Education 2030*. OECD Publishing. https://www.oecd.org/education/2030/

OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en.

OECD (2023), PISA 2022 Results (Volume I): The State of Learning and Equity in Education, PISA, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en.

OECD. (2024). *PISA 2022 Technical Report*. OECD Publishing. https://doi.org/10.1787/01820d6d-en.

Olofsson, G. (2006). *Likvärdig bedömning?: En studie av lärares bedömning av elevarbeten på ett nationellt prov i matematik kurs A* [Fair assessment? A study of teachers' grading of student work on a national test in Mathematics A]. Stockholm: Lärarhögskolan, PRIM.

Palm, T. (2008a). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation, 13*(4), 1–11. https://doi.org/10.7275/0qpc-ws45

Palm, T. (2008b). Interrater reliability in a national assessment of oral mathematical communication. *Nordic Studies in Mathematics Education*, *13*(2), 49–70. https://doi.org/10.7146/nomad.v13i2.148118

Pettersen, A., & Nortvedt, G. A. (2018). Identifying competency demands in mathematical tasks: Recognising what matters. *International Journal of Science and Mathematics Education, 16*, 949–965. https://doi.org/10.1007/s10763-017-9807-5

Pettersen, A., & Braeken, J. (2019). Mathematical competency demands of assessment items: A search for empirical evidence. *International Journal of Science and Mathematics Education, 17*, 405–425. https://doi.org/10.1007/s10763-017-9870-y

Phung, D. V., & Michell, M. J. (2022). Inside teacher assessment decision-making: From judgement gestalts to assessment pathways. *Frontiers in Education, 7*, 830311. https://doi.org/10.3389/feduc.2022.830311

Roos, H., & Bagger, A. (2024). Ethical dilemmas and professional judgment as a pathway to inclusion and equity in mathematics teaching. *ZDM – Mathematics Education, 56*, 435-446. https://doi.org/10.1007/s11858-023-01540-0

Rudner, L. M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation, 2(*4). https://doi.org/10.7275/w4a1-cb66

Skolverket. (2011). *Curriculum for the compulsory school, preschool class and the recreation centre 2011*. Swedish National Agency for Education. https://www.skolverket.se

Skolverket. (2022). *Läroplan för gymnasieskolan 2022 [Curriculum for Upper Secondary Education 2022]*. Swedish National Agency for Education. https://www.skolverket.se

Sollerman, S. (2019). *Kan man räkna med PISA och TIMSS? Relevansen hos internationella storskaliga mätningar i matematik i en nationell kontext* [Can we count on PISA and TIMSS? The relevance of international large-scale assessments in mathematics in a national context] . (Doctoral dissertation, Stockholm University). http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-176740

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1–19. https://doi.org/10.7275/96jp-xz07

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Sage Publications. https://doi.org/10.4135/9781412995627.d5

Sumpter, L. (2013). Themes and interplay of beliefs in mathematical reasoning. *International Journal of Science and Mathematics Education, 11*(5), 1115–1135. https://doi.org/10.1007/s10763-012-9392-6

Svensson, E. (2012). Different ranking approaches defining association and agreement measures of paired ordinal data. *Statistics in Medicine, 31*(26), 3104-3117. https://doi.org/10.1002/sim.5382

von Davier, M., Fishbein, B., & Kennedy, A. (Eds.). (2024). *TIMSS 2023 Technical Report (Methods and Procedures).* Boston College, TIMSS & PIRLS International Study Center. https://timss2023.org/methods

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing.* Routledge. https://doi.org/10.4324/9780415963572