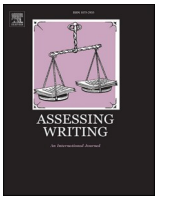
Assessing Writing 57 (2023) 100768

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw



# Beyond literacy and competency – The effects of raters' perceived uncertainty on assessment of writing

Mari Honko [a,*], Reeta Neittaanmäki [a], Scott Jarvis [b], Ari Huhta [a]

[a] Centre for Applied Language Studies, University of Jyväskylä, PL 35, 40014, Finland
[b] Northern Arizona University, 705 S Beaver St, Flagstaff, AZ 86011, USA

ARTICLE INFO

*Keywords:*
Rater behavior
Uncertainty
Confidence
Rating quality
Language assessment
Assessing writing

ABSTRACT

This study investigated how common raters' experiences of uncertainty in high-stakes testing are before, during, and after the rating of writing performances, what these feelings of uncertainty are, and what reasons might underlie such feelings. We also examined if uncertainty was related to raters' rating experience or to the quality of their ratings. The data were gathered from the writing raters (n = 23) in the Finnish National Certificates of Proficiency, a standardized Finnish high-stakes language examination. The data comprise 12,118 ratings as well as raters' survey responses and notes during rating sessions. The responses were analyzed by using thematic content analysis and the ratings by descriptive statistics and Many-Facets Rasch analyses. The results show that uncertainty is variable and individual, and that even highly experienced raters can feel unsure about (some of) their ratings. However, uncertainty was not related to rating quality (consistency or severity/leniency). Nor did uncertainty diminish with growing experience. Uncertainty during actual ratings was typically associated with the characteristics of the rated performances but also with other, more general and rater-related or situational factors. Other reasons external to the rating session were also identified for uncertainty, such as those related to the raters themselves. An analysis of the double-rated performances shows that although similar performance-related reasons seemed to cause uncertainty for different raters, their uncertainty was largely associated with different test-takers' performances. While uncertainty can be seen as a natural part of holistic ratings in high-stakes tests, the study shows that even if uncertainty is not associated with the quality of ratings, we should constantly seek ways to address uncertainty in language testing, for example by developing rating scales and rater training. This may make raters' work easier and less burdensome.

## 1. Introduction

Rater behavior has been studied extensively (see, e.g., Reed & Cohen, 2001; Lumley, 2005; Pill & Smart, 2021), and individual raters have been found to differ in terms of consistency (Davies et al., 1999; Van Moere, 2013), severity (Eckes, 2005; Huhta et al., 2014; Schaefer, 2008; Wind et al., 2021), what they prioritize (Eckes, 2008, 2017; Wind et al., 2019), and the strategies they rely on when assessing high- vs. low-proficiency performances (Kuiken & Vedder, 2014). Even though the causes and effects of rater

* Corresponding author.
*E-mail addresses:* mari.h.honko@jyu.fi (M. Honko), reeta.m.neittaanmaki@jyu.fi (R. Neittaanmäki), scott.jarvis@nau.edu (S. Jarvis), ari.huhta@jyu.fi (A. Huhta).

https://doi.org/10.1016/j.asw.2023.100768
Received 30 June 2022; Received in revised form 7 July 2023; Accepted 9 July 2023
Available online 25 July 2023

uncertainty have rarely been empirically examined in language assessment, researchers in this and related fields (e.g., communication sciences and disorders, second language acquisition) are indeed aware of the importance of rater uncertainty and its potential effect on rating consistency (e.g., Davies et al., 1999, p. 91; Fitzpatrick & Thwaites, 2020; Youn, 2018). Recommendations for increasing raters' confidence have also been given (e.g., Alderson et al., 1995, p. 128; Gorsuch & Griffee, 2018, p. 104). Self-reported confidence evaluations are increasingly being built into assessment protocols and the research designs of studies on language ability (Barkaoui, 2011; Bogorevich, 2018; Borders et al., 2021; Bosshardt et al., 2016) but the data these studies have generated on rater confidence and uncertainty have rarely been systematically analyzed in published language assessment research. The present study aims to fill this research gap and examines rater uncertainty in the assessment of writing using multi-method analysis and empirical data. The data come from the National Certificates of Language Proficiency, a high-stakes examination in Finland with over 9000 participants each year. Our research questions are the following:

Q1 Do writing raters experience uncertainty before, during or after assessment?

a. How common are the experiences of uncertainty?
b. What reasons for uncertainty are raters aware of before engaging in the assessment process?
c. What sources of uncertainty do raters report during and after the assessment?

Q2 To what extent does rater uncertainty reflect raters' level of experience?

Q3 To what extent is the uncertainty experienced by raters related to the quality of their assessments measured as

a. consistency?
b. severity vs. leniency?

Q4 What do writing raters think the test organizer needs to do in order to address potential problems related to rater uncertainty?

In this paper, we view rater uncertainty as the rater's lack of confidence in assigning a score to a particular sample of language performance. This is a phenomenon that is scalar rather than binary; it will rarely be the case that a rater is either completely confident or completely uncertain about a particular rating. Most raters in most situations will fall somewhere between these two extremes. Nevertheless, there is a point along the continuum where confidence becomes uncertainty, and in this study we investigate how often raters face uncertainty, what causes raters to find themselves on the uncertainty side of the continuum, as well as what effects—if any—uncertainty has on the quality of their ratings.

## 2. Rater uncertainty

### 2.1. Uncertainty in decision making

Rater uncertainty has received special attention in medical and legal research, where the accuracy of ratings (alternatively referred to as diagnoses, judgments, decisions, rulings, and verdicts) entails exceedingly high stakes for the people whose conditions or behaviors are being evaluated (e.g., Bhise et al., 2018; Posner, 2008). An assumption that underlies many of these studies is that uncertainty can result in errors, delays, and excessive use of resources in decision making (Bhise et al., 2018; Posner, 2008) – all of which can of course have serious consequences for people's lives and well-being. It is helpful to view rater uncertainty in language assessment as uncertainty in decision making (cf. Cumming et al., 2002, p. 88). Uncertainty in decision making has been the focus of a good deal of research in psychology since at least the 1970s (e.g., Milburn & Billings, 1976). In a recent review of the research in this area, Anderson et al. (2019) pointed out that uncertainty is ubiquitous in daily human life and that it has been investigated from several disciplinary perspectives using different terminology and different theoretical and methodological frameworks. Situations where people engage in subjective decision making—such as when assessing performance skills such as speech and writing—are especially fertile ground for uncertainty (see also Pill & Smart, 2021; Tarnanen, 2014).

The findings of the different strands of research on uncertainty in decision making indicate, among other things, that there are individual differences in raters' propensity for uncertainty, and that uncertainty can dampen as well as amplify the affective dimensions of decision making. In some cases, it appears that uncertainty increases people's reliance on affect (e.g., the rater's emotional state or the esthetic characteristics of the rated performance) when making decisions (Faraji-Rad & Pham, 2013).

### 2.2. Uncertainty in language assessment

When rater uncertainty is addressed in the literature on language assessment, it is often brought up as an incidental observation of one of the many factors that might affect ratings. For example, in Ahola's (2016) study, the interviewed professional raters reported that their confidence increased with experience, but that uncertainty remained an inherent part of assessment. Tarnanen (2002) similarly reported that even an experienced rater can be left with feelings of uncertainty. Experience nevertheless allows the raters to develop greater awareness of their attitudes and expectations, as well as tolerance for uncertainty (Ahola, 2016).

According to our own experience as language testers and the qualitative studies of Tarnanen (2002) and Ahola (2016), uncertainty can be related, for example, to the rater's background and experiences, skills, personality and behavioral or emotional predispositions (rater-related factors), the quality of specific performances being assessed (performance related factors), assessment processes, criteria and instructions (system-related factors) or to other factors, such as distractions and disruptions or the feeling of being rushed during

assessment. The former factors are more clearly internal to the assessment, the latter more clearly external. However, the relationship between these factors should not be assumed to be clear-cut, but complex, because raters' ability to tolerate for example disturbances probably varies across individuals. Therefore, the same external factors cannot be assumed to cause uncertainty in all raters or in the same way since uncertainty in assessment is realized at the individual level. However, it is possible that some factors, such as certain types of performance or assessment environments, would cause uncertainty systematically.

Besides the above studies in language assessment, only a handful of other studies in the language sciences shed light on the causes and effects of rater uncertainty. Regarding its causes, studies on speech pathology and communication disorders have found that problems related to data quality (e.g., video quality, camera angle, image stability) are the main sources of uncertainty in clinical workers' ratings of patients' vocal tract problems (Borders et al., 2021). However, the type of data also matters. Canino (2001) found that raters were less confident in rating the emotional expressions of participants with brain damage than they were in rating comparable data produced by normal controls. Canino also found that the raters' confidence was affected by the participants' accuracy and emotional intensity in the expressions they produced.

In research on language assessment, Bogorevich (2018) found that familiarity with participants' accents improves rater confidence when assessing speaking performance. Hsu (2012) similarly found that unfamiliarity with phrases or structures produced by speakers of Indian English resulted in uncertainty among raters of the IELTS speaking test. More specifically, the raters were not sure whether the unfamiliar phrases and structures were conventional in Indian English or whether they were vestiges of learner language. Hsu's study also suggested that raters had their own individual interpretations of what constituted L2 speaking competency (p. 83). In a study on the use of think-aloud protocols during essay rating, Barkaoui (2011) reported that the use of think-aloud protocols lowered the confidence of several of the raters because it caused them to second-guess themselves (p. 65).

The only study in the language sciences we are aware of that has examined the effects of rater uncertainty on the quality of ratings is the one by Bosshardt et al. (2016), who investigated the ratings of 170 speech language pathologists (SLPs) who were asked to use a 10-point scale to rate the severity of stuttering in seven pre-school age children representing the following seven L1 backgrounds: Danish, English, French, German, Greek, Italian, and Persian. Each rater was proficient in at least one of these languages, but the raters were also asked to rate stuttering severity in children whose languages they were not proficient in. The raters also indicated their level of confidence using a 10-point scale. The results showed that the raters' proficiency in the children's languages did not significantly affect their ratings, nor did the raters' self-reported levels of confidence or their years of experience as SLPs affect their ratings. These findings therefore suggest that rater uncertainty does not necessarily compromise rating quality (e.g. consistency, severity/leniency). All these studies examined the assessment of speaking, however, so the results may not generalize to writing. The Bosshardt et al. study also differed from the present study in that stuttering is a conceptually simpler and observationally more superficial phenomenon than writing proficiency. There is a lack of similar research on writing, and the goal of our research is to address this research gap.

We conclude this section by mentioning some general factors that are likely to affect rater uncertainty. Pill & Smart (2021) have noted individual differences among language assessment professionals concerning how natural and pleasant they find the experience of rating and how talented they consider themselves to be as raters. Raters are presumably the most confident in their own rating ability when they carry out rating voluntarily. Rater uncertainty can be decreased when testing protocols include quality control measures such as rater training, clear rating criteria, and a carefully planned assessment process (Ahola, 2016, 2022). By contrast, rater uncertainty can be heightened in high stakes situations where raters know that the test scores will have major consequences for the examinees (Ahola, 2016).

## 3. Present study

### *3.1. Rating in the national certificates*

The National Certificates of Language Proficiency (NCLP) is an official language examination in Finland. It is supervised by the Finnish National Agency for Education (FNAE) but the responsibility for designing the tests and ensuring the quality of rating lies with the University of Jyväskylä. The NCLP measures adult learners' functional general language proficiency in the official languages of Finland – Finnish and Swedish – and in seven other languages. Three levels of examinations are available: basic, intermediate, and advanced. The NCLP is a high-stakes examination used, for example, for demonstrating the required level of language proficiency for Finnish citizenship and for various study or work purposes. The intermediate level examination of Finnish has the largest number of participants and most of the examinees take it as part of their naturalization process (Neittaanmäki & Hirvelä, 2014). This examination has also been audited by the Association of Language Testers in Europe.

Assessment in the NCLP is criterion-referenced, and the same rating scales and standardized rating procedures are applied across all languages and administrations of the examinations. Examinees' writing performances are rated with reference to the NCLP criteria (FNAE, 2011). In the intermediate level examination, three grades (levels) are given: below 3 (the performance does not meet the criteria of the proficiency level 3), 3, and 4. These grades correspond to the following Common European Framework (Council of Europe, 2001) levels: below B1 (does not meet the B1 level criteria), B1 and B2.

The writing test comprises three tasks. Examinee performances are rated either remotely or in on-site rating sessions. In the NCLP, at least half of the writing performances are double-rated (66%, in this study). All ratings are linked through double-marking so that rating quality can be monitored. In addition, a number of performances are selected for further checks based on the statistical analyses of the ratings.

The NCLP raters are required to have university level education in Finnish and they mostly work as L2 Finnish teachers. They have to successfully complete an initial rater training, after which the FNAE grants them a license to rate NCLP performances. The raters also

participate in training events before each new examination round. Rating quality is addressed by monitoring raters' use of the rating scale and their consistency and severity. This is done through Multi-Facet Rasch analyses using the Facets software (Linacre, 2021). Whenever inconsistent or too severe/lenient ratings are discovered, those ratings are scrutinized, and feedback is given to the rater.

### 3.2. Materials and methods

#### 3.2.1. Research procedure and data

The study was conducted in 2021–22 as part of the Finnish intermediate examination. 23 raters of writing participated in the research. They used an on-line system (i.e., rated remotely) except for 13 raters who participated in joint, on-site assessment sessions. All raters completed training specific to the particular examination before starting their work, including the rating of practice samples. Participation in the study was voluntary; the participants were given feedback and a modest additional payment.

The data comprise three parts: (1) raters' responses to on-line surveys (before, in the middle of, and after the examination round), (2) raters' notes during the rating sessions regarding particular performances and tasks, and (3) ratings and the statistical indices related to those ratings. Information about the raters' experience in working for the NCLP was also available. The survey instrument (Webropol surveys and Excel note-taking sheet) developed for the study are available online (Honko et al., 2023).

The rater surveys were administered through the on-line system called Webropol. Both selected-response items and free response questions were used.

The pre-rating survey used Likert-type items that elicited raters' estimations of how difficult, in general, they considered rating to be and how confident they generally were in their ratings. The raters were also asked about their confidence when preparing for the upcoming ratings. In addition, they were asked to write down reasons for any feelings of uncertainty.

The second survey collected raters' experiences about the rating process (uncertainty during a particular session; reasons for uncertainty) and conditions during either remote or on-site sessions. Raters completed this survey after every extended rating period. The final number of surveys filled out by individual raters varied from one to six. The main reason for the variation is that the raters could take breaks from their work in different ways. A few of them also seemed not to follow the instructions precisely, and they probably did not complete as many surveys as they had sessions. The raters were asked to use Likert-type survey items to estimate their confidence and the suitability of the conditions for rating and to answer three selected-response items about their feelings of uncertainty related to their (1) interpretation of the criteria, (2) consistency, and (3) severity/leniency. They could also add to this list of uncertainties if other types of uncertainty emerged.

The final survey encouraged the raters to reflect on how NLCP raters' confidence in their ratings could be increased and their general views on participating in the study.

The raters produced notes about the assessments they felt uncertain about if they remained uncertain after giving the rating. In addition, the raters were asked to write reasons for their uncertainty and measures they took to address such uncertainties. They could choose from a menu of possible reasons for uncertainty related to the test-taker's performance such as (1) borderline case between two proficiency levels, (2) uneven performance, (3) uncertainty about the degree to which the examinee fulfilled the task requirements, (4) brevity of performance, and (5) some other reason (to be elaborated by the rater).

The rating data for writing came from 2496 examinees (4144 rated performances due to double ratings). Since each examinee completed several (usually three) writing tasks during the examination, the rating data included 12,118 data points. In this article, we use (examinee) performance comprising their performance across all the writing tasks in the particular writing test as the main unit of interest. This is also justified by the fact that when double-rating, the raters assess the examinee's performance on all (three) tasks, not just on one of them.

#### 3.2.2. Analyses

The free response data were analyzed by using thematic content analysis (Braun & Clarke, 2006). Each rater's (sub)themes were then grouped into larger thematic categories.

The quantitative data were analyzed by using descriptive statistics. In addition, the associations between the variables of interest were investigated by using scatterplots and Spearman rank-order correlations due to the ordinal nature of the Likert scale-type responses or their non-normal distribution.

To address Research Question 3, we estimated raters' severity and within-rater consistency by applying a Many-Facet Rasch Measurement model (MFRM) (Linacre, 1994; Eckes, 2011) using Facets software (Linacre, 2021). The MFRM is widely used and "considered one of the most useful validation tools in studies of rater effects" in the field of language assessment (Aryadoust et al., 2021 p. 6) since it allows the simultaneous estimation of the key facets of rating (learner ability, rater severity, task difficulty) and their calibration on the same linear (logit) scale. In the MFRM, rater severity was measured in logit values and rater consistency as rater infit mean-square values. Bias analysis was used to examine whether raters were more lenient/severe when rating an entire test if they were uncertain about at least one of the tasks within that test.

## 4. Findings

### 4.1. Raters' perceptions of rating

#### 4.1.1. Prior to the rating sessions

Rater confidence/uncertainty and rating facility/difficulty were examined through the initial survey administered to the raters

prior to the rating sessions. Only 40% of the writing raters reported that they were, in general, always or nearly always confident in their ratings (Fig. 1). With only one exception, all raters reported that they were generally at least somewhat more often confident than uncertain regarding the ratings they give, and uncertainty was not a prevailing concern for any of them before the rating sessions.

Raters' feelings of confidence were in line with how easy or difficult they generally considered rating to be. Confidence/uncertainty and facility/difficulty were measured with similar Likert-type survey items, and a Spearman rank-order correlation showed that the association between the two variables was moderate/strong and significant ($r_S$=0.653, $p < 0.001$).

The initial survey additionally asked the raters how confident they were while preparing for the particular rating round, and what factors they thought affected their confidence. In their open responses, the raters reported that rater training and an ability to achieve a sufficient level of consensus in the norming process during training were especially important for enhancing confidence. They also indicated that uncertainty would be more likely if they had differing perspectives on the training samples, when they might rely too much on their own subjective views, if they were insufficiently trained, and if they had not rated for a long time. Lack of experience was also a cause for concern. However, the raters' responses to selected-response items showed that uncertainty was not a strong or even a top concern for them: 79 % of the raters reported that they were confident or fairly confident when beginning the rating process (Fig. 2).

Additionally, some reported that the challenges of rating and the feelings of uncertainty were not exclusively negative; instead, these increase the meaningfulness of their work and support the development of one's professional skills. Uncertainty can, for example, help raters to reflect on their assessment behavior, look for ways to maintain quality in their assessment, and devote more time to rating a particular text when needed. In addition, challenging oneself can increase the appreciation of one's own expertize as a rater.

#### *4.1.2. During the rating sessions*

While assessing, the raters wrote notes on the parts of the tests that they were uncertain about. Notes were task-specific and were made only when the rater *remained uncertain* after assessing. The raters indicated that they were confident in their ratings of 87 % (2174) of the 2496 written performances. These performances did not include any task-specific ratings marked as uncertain. (Recall that the writing tests include three tasks and that here performance refers to the examinee's performance across all three tasks). When the rater experienced uncertainty while rating a performance (approximately 13 % of the performances), the rater's uncertainty was usually confined to a single task within that test – i.e., just one of the three texts written by the same writer (n = 155, approximately 6 % of the writing performances). In some cases, a rater's uncertainty extended to two tasks within an examinee's performance (n = 62, approximately 3 % of the performances). However, raters experienced slightly more uncertainty in rating all three tasks produced by the same writer (n = 108, about 4 % of the performances).

Fig. 3 illustrates rater-specific variability in relation to the amount of uncertainty each rater experienced during rating. The figure shows the proportion of performances rated by each rater that included at least one task the rater was uncertain about. For just over a third of the raters (n = 8), the number of performances consisting of one or more uncertain tasks was less than 5 % of all performances. For just under a third of the raters (n = 7), uncertain ratings were limited to 5–9 % of the performances. Three raters were uncertain about 10–14 % of the performances they rated, and three additional raters were uncertain about 15–19 %. Only one rater experienced uncertainty more often than this (in 20 % or more of the performances rated).

The analysis described earlier revealed differences between raters in the amount of uncertainty during the rating sessions, but it did not consider whether the raters' uncertainty related to the same performances. To address this latter question, we examined performances that were rated by two raters. The writing data include 1644 peformances (66 % of the performances) that were double-rated.

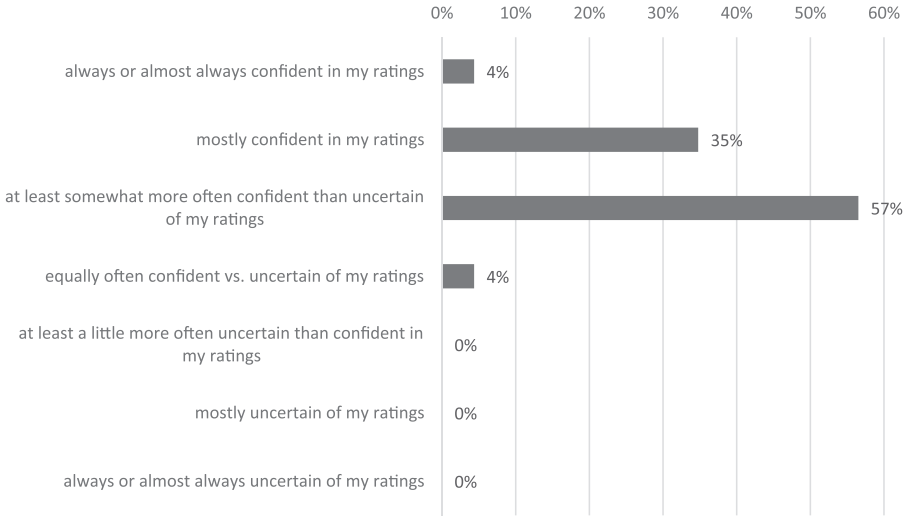



**Fig. 1.** General rater confidence according to the survey conducted before the beginning of the assessment ("In general, how confident are you in your writing ratings?").

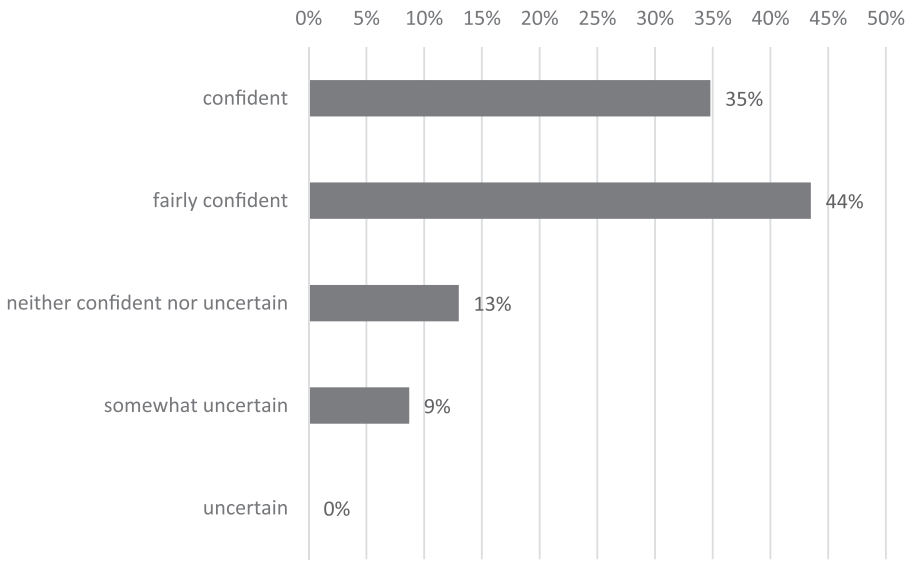



**Fig. 2.** Raters' confidence prior to the rating sessions ("How are you feeling before starting the x month's assessment?").

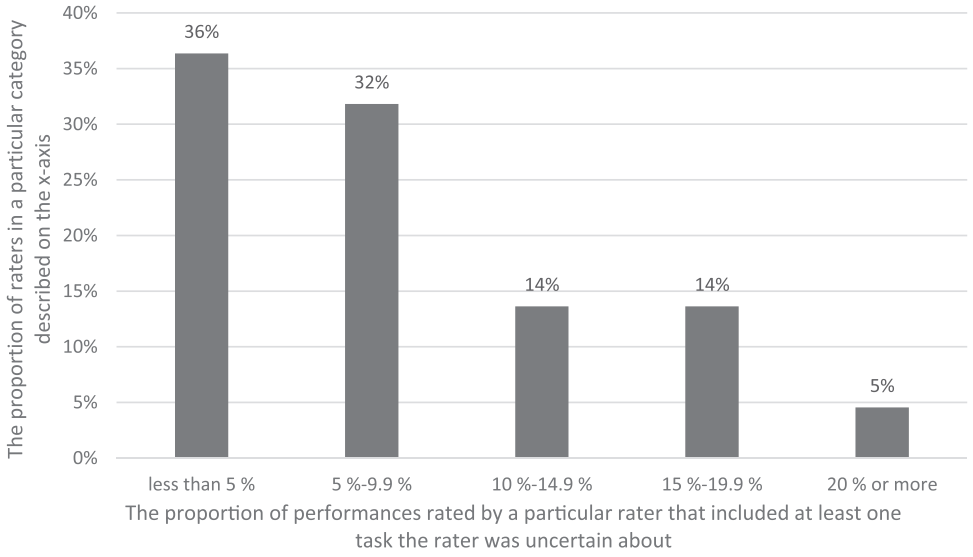



**Fig. 3.** Rater-specific variability in relation to the amount of uncertainty.

The majority of these (n = 1388, 84 % of the double-rated performances) did not cause uncertainty for any rater. A total of 256 performances received uncertain ratings, which was 16 % of all double-rated performances. However, very few of these received uncertain ratings from both raters. There were only 13 performances that both raters felt uncertain about, and this represents only 5 % of the uncertain double-rated performances and 0.8 % of all double-rated performances (Appendix A).

#### *4.1.3. Immediately after the rating session*

The raters' responses on the surveys completed immediately after each rating session indicate that they considered the rating conditions – e.g., the surroundings, lack of distractions, level of alertness – to be mostly good or very good. However, the on-site group gave somewhat higher ratings to their rating conditions than the online group. The raters' responses also show that their perceptions of uncertainty were somewhat similar to what could be found in their notes made before and during the rating sessions (Figs. 3 and 4), but some of their perceptions also differ. The session surveys underscore that there were individual differences in the degree of uncertainty that different raters experienced, as can also be seen in their session notes and their responses on the pre-rating survey. On the other hand, feelings of uncertainty appear to be somewhat more accentuated in the raters' post-session reflections. It appears that at least some raters may have felt that their own session-specific uncertainty was greater than what their session notes indicate (see Fig. 4). Confirmation of this interpretation would require a more detailed analysis of the session-specific data, but it is supported by our

preliminary observation that there was no clear change or substantial variability in raters' session-specific uncertainty after the first session: uncertain ratings were slightly more numerous in the first session than in later sessions, but there were no observable systematic differences across the remaining sessions. Thus, session-specific fluctuations do not explain why, for example, only 18 % of the responses in the post-session survey indicated that the raters were always or almost always confident in their ratings (Fig. 4), whereas the notes they wrote during the rating indicated that more than a third (36 %) of them were rarely uncertain of their ratings: they were certain with all tasks in at least 95 % of the performances assessed (Fig. 3).

### 4.2. Sources of uncertainty described by the raters

#### 4.2.1. Prior to the rating sessions

The pre-rating survey asked raters to describe what they believed had caused uncertainty in their ratings in the past. Their responses were thematically organized in relation to all the ideas expressed. Six main themes emerged regarding the causes of uncertainty: (1) the characteristics of the test performance being rated, (2) uncertainty about one's own approach to rating, (3) mental or physical fatigue, (4) matters relating to the task, criteria, and training, (5) distractions, and (6) remote assessment.

The most frequently reported causes of uncertainty (14 raters, or 61 % of the raters) involved the characteristics of the test performance being assessed. The most often mentioned characteristics were borderline cases (example 1), uneven performance (example 2) and performance ambiguity such as poor handwriting (example 3). Other characteristics included task fulfillment (example 4), an insufficient sample (example 5) and a mismatch between the examinee's linguistic expressions and other indicators of their competence (example 6). The original excerpts reflecting these themes were written in Finnish and are provided in Appendix B.

(1) "borderline cases of course always make you wonder"
(2) "uneven writing quality, where one part of the text represents higher writing quality than the other parts"
(3) "poor handwriting"
(4) "Uncertainty is mostly caused by texts that are off topic but well written."
(5) "Sometimes it seems that there is too little evidence to say anything with certainty about their language abilities."
(6) "Also difficult are writers who might be good writers in other languages but whose Finnish skills are not yet at a sufficiently high level."

In addition to characteristics of test performance, nearly as many raters expressed uncertainty about their own approach to rating, or their perceptions of themselves as raters (12 mentions, about 52% of the raters, see also Vaughan, 1991 and Eckes, 2008 about rater style). The factors related to these perceptions mentioned by multiple raters include uncertainty about how to weight the criteria (example 7), uncertainty about how to apply their familiarity with L2 Finnish (concern about understanding learner language too well/easily) in the reading of texts written by learners of Finnish (example 8), and lack of rating experience (example 9). Some raters also mentioned mental or physical fatigue and being stressed about being in a hurry (examples 10 and 11).

(7) "deliberating about what to emphasize in one's rating and how to bring it all together into a holistic assessment"

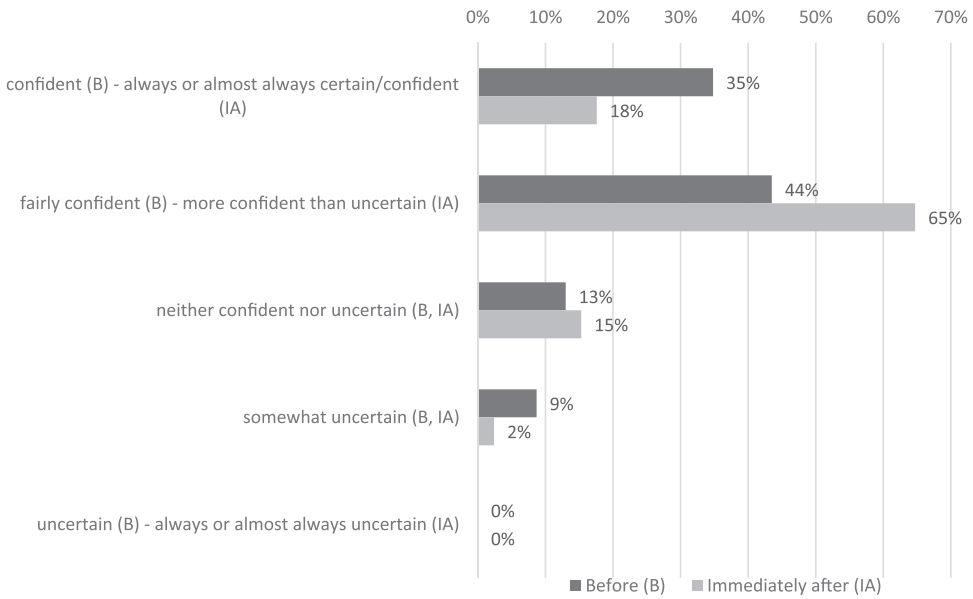



**Fig. 4.** Raters' confidence before and immediately after the rating sessions.

(8) "the so-called L2 Finnish teacher dilemma, or whether a layperson would understand this"
(9) "Whether it has been a long time since the previous assessment round."
(10) "Sometimes things in my own life and a sore neck and shoulder bother me and make it difficult to rate texts, my working posture is sometimes uncomfortable."
(11) "Feeling rushed. Fatigue."

Finally, 8 raters (35 %) pointed to uncertainty arising from the test design, including the task instructions and prompts, scoring criteria, and rater training (examples 12 and 13). Here, uncertainty was described as occurring inter alia in situations where the instructions for a particular test task were ambiguous or where the scoring criteria or rater training did not clarify how a given performance should be rated.

(12) "Sometimes the instructions are not unambiguous, and neither are the criteria."
(13) "if it feels like the criteria are not helpful (these could sometimes be more elaborate)"

Other potential causes of uncertainty outside the aforementioned categories were mentioned only once. These included unspecified external distractions and the number of test performances being assessed.

*4.2.2. During the rating sessions*

Raters' notes during the rating sessions show that uncertainty occurred by far most often (56 % of all mentions) when an examinee's performance was on the borderline between two proficiency levels. An inspection of the data shows that uncertainty in borderline cases may be more typical at the threshold for level 3, which is a more meaningful cut-off for decision making than for level 4, but further research is needed to confirm this. Raters' notes also show that uneven test performance caused uncertainty quite often. In comparison, uncertainty associated with task fulfillment and insufficient samples was less frequent.

The information presented in Fig. 5 comes from raters' responses to the selected-response items in the survey during the rating sessions. It is useful to take a closer look at the causes of uncertainty they provided under the option "other reason", such as level of alertness, (in)consistency, or distractions. The raters could choose more than one option, but in over 70% of the cases, raters chose only one of the ready-made options. This means that causes of uncertainty other than those described in the options were not reported by the raters very often during the rating sessions. Moreover, some of the "other reasons" of uncertainty given by the raters were paraphrases of the existing options (e.g., problems related to task fulfillment). Half (n = 41) of the other causes of uncertainty pertained to unclear handwriting and its impact on comprehensibility and ease of reading (example 14). Further causes of uncertainty mentioned by multiple raters included confusing or unclear content (14 mentions, example 15), deficiencies in the control of a particular criterion, such as language structures (example 16) and a large number of spelling or punctuation errors (n = 6).

(14) "the person writes a lot, but I struggle because of handwriting that is difficult to read, which disrupts my reading rhythm".
(15) "The intended meaning emerges but in a way that is difficult to follow".
(16) "There is quite a lot of content but the language structures are really awkward.".

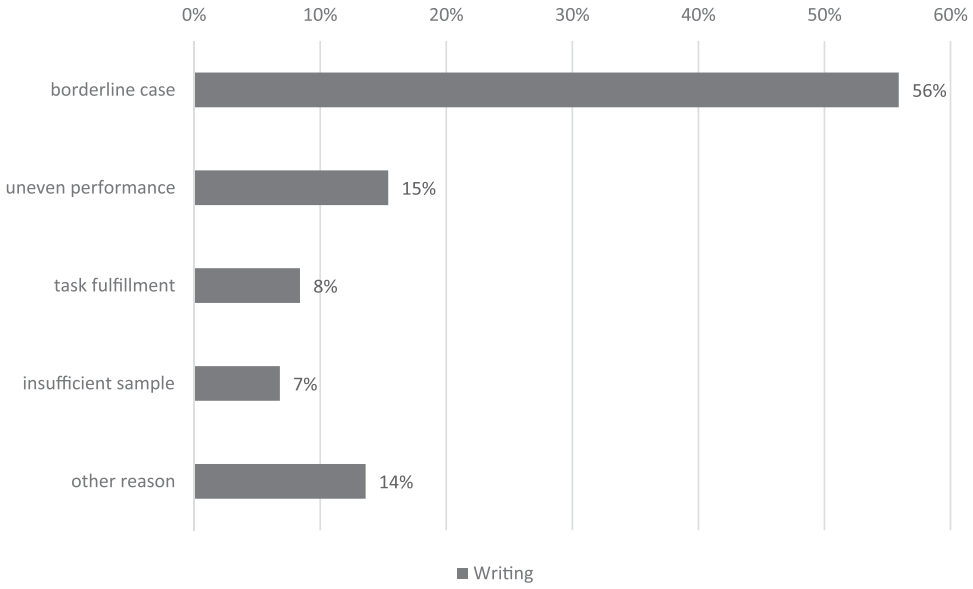



**Fig. 5.** Causes of uncertainty reported by the raters during the rating sessions.

A few other isolated causes of uncertainty were mentioned, and these focused mainly on factors related to test performance (e.g., redundancy and the impact of erroneous facts on comprehensibility). The raters did not provide any comments on the assessment conditions, and there were only three comments pertaining to raters themselves: one rater related uncertainty to familiarity with speakers of L2 Finnish, speculating that she might understand this population much better than Finnish L1 speakers who do not have her background; another rater related uncertainty to the rater's state of alertness, and a third to the consequences of the rating for the examinee's future.

#### *4.2.3. Immediately after the rating sessions*

The third type of survey collected both quantitative and qualitative data on the causes of uncertainty immediately after each rating session. The raters' responses to the selected-response items on this survey show that, during most sessions, raters experienced at most occasional uncertainty about (a) how to interpret the scoring criteria (7 % more than occasionally), (b) their own severity/leniency (13 % more than occasionally), or (c) their consistency (17 % more than occasionally). Despite the relatively low amount of uncertainty reported and differences in survey questionnaire design, these results nevertheless show that the raters identified substantially more causes of uncertainty after the sessions than they did during them. Raters were not required to answer the open-ended question "If you experienced uncertainty during this session, what caused it?," but this question generated 64 responses, with some raters responding to it more than once across sessions. In these responses, raters referred to a wider range of causes of uncertainty than they did in the notes they made during the sessions.

A content analysis showed that almost half of the mentioned causes related to such performance characteristics as handwriting, borderline cases, uneven performance or sample size, etc. (a total of 34 mentions), and a few referred to how the performance characteristics related to each other or the interpretation or weighting of the scoring criteria. Other responses (48 mentions) dealt with factors related to their perceptions of themselves as raters ($n = 15$), physical distractions (fatigue, alertness, brain fog, pain, new glasses, $n = 7$), lack of experience or difficulty getting started ($n = 5$), responsibility to examinees and concerns about impartiality ($n = 3$), and concerns about whether there was sufficient time ($n = 2$). Additional causes of uncertainty included familiarity with L2 Finnish, comparisons of different test performances, and the overall number of examinees receiving poor ratings. The effects of the rating environment were mentioned once, and three additional responses mentioned aspects related to participating in the study. However, none of these latter responses suggested that study participation per se increased rater uncertainty.

### *4.3. The relationship between rater experience and rater uncertainty*

The relationship between rater experience and rater uncertainty was examined by correlating the raters' years of experience as NCLP raters with the levels of uncertainty they reported. Their years of experience varied from just over one year to almost 28 years (min 1.4 and max 27.8, mean 15.0, median 10.8, standard deviation 9.2). All raters had been actively involved in NCLP assessment, but they did not all have recent experience in rating writing. Uncertainty was examined in relation to each of the following four variables: overall rater confidence before the rating sessions (Fig. 1), rater confidence at the start of the rating sessions (Fig. 2), the proportion of tests marked as uncertain during the rating sessions (Fig. 3), and the amount of uncertainty reported after the rating sessions (Fig. 4). A series of Spearman's rank-order correlation analyses showed that rater experience was not correlated with the four measures of rater uncertainty ($r_S = -0.103$ to $0.351$, $p > 0.05$). This was confirmed through an examination of the relevant scatterplots.

### *4.4. The relationship between rater uncertainty and the quality of their assessment*

Rater severity and consistency were analyzed using the Many-Facet Rasch Model. The overall data-model fit was satisfactory (only 2.9 % of absolute values of standardized residuals were equal or greater than 2 %, and 1.0 % equal or greater than 3). The separation statistics (e.g., reliability, separation) indicating the reproducibility of the measures were also within acceptable limits. Our focus is on the raters, but Appendix C summarizes the analysis of the other facets (examinees and tasks). The rater facet was constrained with a mean rater severity of 0 and ranged from − 1.56–1.52 logits (the lower the measure, the more severe the rater). The standard errors for rater severity were small (ranging from 0.09 to 0.17), indicating high measurement precision. The mean infit and outfit mean-square values were 0.98 (SD = 0.12) and 0.96 (SD = 0.23), respectively. The reliability of the rater measures reached values as high as 0.98, indicating large differences in the rater severity. Infit mean-square values (consistency) varied between 0.77 and 1.28. Considering values between 0.70 and 1.30 acceptable, we can conclude that all the raters were sufficiently consistent.

These MFRM results support our conclusion that rater severity and consistency indices were sufficiently valid and reliable for the analyses required to address Research Question 3. Next, we examined the relationship between perceived uncertainty and rater consistency. Because there was little variance in the response variable representing consistency, a statistical relationship cannot be determined between uncertainty and consistency. Importantly, although all raters reported at least occasional uncertainty, this appears not to have decreased the consistency of their ratings. Although it is not meaningful to investigate statistically the connection between uncertainty and consistency due to the raters' (high) consistency and the small variation in the data, the result regarding consistency is still an important part of the research. Namely, if it could be generalized that rater uncertainty causes inconsistency, the most uncertain raters in our data would have been more inconsistent than the others. However, this did not happen. It can be concluded that a reasonable amount of uncertainty does not increase inconsistency in writing assessment if there are no other quality problems in the assessment.

It is possible that a rater's behavior (i.e., approach, strategies, tendencies) changes while rating performances that the rater is uncertain about. We therefore examined the relationship between perceived uncertainty (4 rater-specific variables) and rater-specific

levels of severity. A series of Spearman's rank-order correlation tests showed that rater severity was not statistically correlated with any of the four measures of rater uncertainty ($r_S$=−0.085 to 0.225, $p > 0.05$). Fig. 6 illustrates the lack of a linear relationship between the two variables: level of uncertainty and severity. This observation further supports the conclusion that raters' levels of uncertainty did not affect the quality of their ratings. In addition, the results of bias analyses showed that the raters' severity did not vary significantly with respect to their confidence about the performances they were rating.

### 4.5. Raters' perception of support needed to reduce uncertainty

The final survey asked raters to describe the types of support they would like to receive from the organization responsible for administering the language tests. Half of the raters (n = 11) expressed a need for ongoing rater training (example 17). The raters felt that it is less important to offer new types of support than it is to hold rater training sessions in connection with each round of assessment, in a way that allows raters to get together to review the test tasks, scoring criteria, and individual performance samples. The raters found the rating conditions to be quite good or very good during the study, but several (n = 7) hoped that the rating conditions would also be controlled and monitored in the future.

(17) "Joint rater training is absolutely essential and helps immensely every time."
(18) "maybe returning to in-person work where we are in close physical proximity to our colleagues will help because there are more distractions in the assessment environment at home"

The raters expressed a particular desire for focused (on-site) rating sessions rather than moving entirely to remote rating (example 18). Some (n = 5) called for further development of the assessment process or guidelines, such as by offering more opportunities to discuss the rating of uneven, challenging, and complex task performances. Two raters also expressed the need for more rater feedback, including personal feedback after every rating session. Additional wishes included further development of test tasks (e.g., reducing ambiguity), different modes of rater training (lively remote training, but also on-site training), more comprehensive rater training samples, importance of double-rating and careful rater selection. It should be noted, however, that 6 raters also mentioned that they are pleased with things as they are.

## 5. Discussion and conclusions

Rater uncertainty has been assumed to decrease, and the means to deal with it to increase, with experience (e.g., Tarnanen, 2002; Ahola, 2016), but raters' uncertainty related to high-stakes language testing is rarely discussed. This study shows that the majority of the investigated raters were, in general, confident or fairly confident in the ratings given and in themselves as raters. Nevertheless, even trained raters were at least occasionally uncertain about specific ratings. Both the surveys and raters' notes showed that rating uncertainty caused especially by task-related and subjective factors can be found even among experienced raters. External factors, such as noise or interruptions, appeared only relatively rarely among the reasons for the uncertainty. This can be explained by the fact that distractions were rare, but the raters might also have had good strategies for dealing with distractions that arose during assessment, so

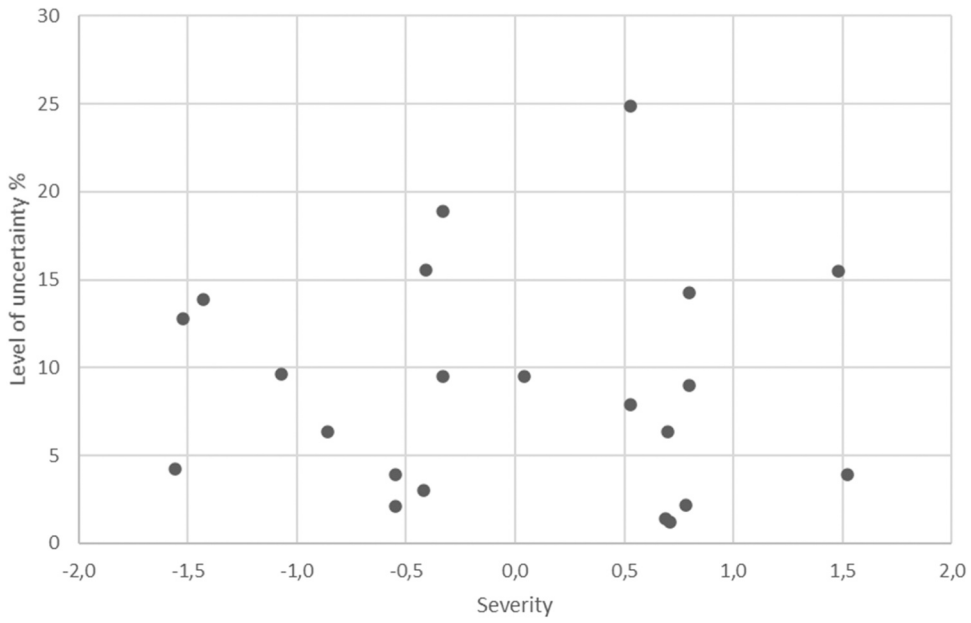



**Fig. 6.** Raters' severity/leniency compared to rater uncertainty.

that these had no direct effect on their rating certainty.

Importantly, we found large individual differences across raters regarding the frequency of their feelings of uncertainty. We also found that raters rarely feel uncertain about the same performances and tasks, even when the assessments were done at the same time and under the same conditions in on-site rating sessions. This suggests that uncertainty, at least in a controlled test environment, is highly individual and that raters, for example, do not necessarily agree about which test performances are borderline cases and, therefore, a reason to be uncertain. Raters reflect on their own assessment in different ways—for example, not everyone ponders their certainty/uncertainty equally deeply or views their own uncertainty in the same way. The handling of external assessment conditions is also individual in that the same distractions do not affect raters' assessment certainty in the same way. A fourth important finding of the study was a confirmation of the observation by Bosshardt et al. (2016) that rater uncertainty does not appear to affect rating quality. In our analyses, uncertainty was not found to vary as a function of rater consistency or severity/leniency (cf., e.g., Youn, 2018).

According to the raters, limited rating experience and long intervals between rating opportunities increase rater uncertainty. However, as also Tarnanen (2002) and Ahola (2016) stated, uncertainty is not solely or even primarily tied to the rater's experience; we found that even highly experienced raters felt uncertain about several performances. Importantly, our results showed that rater experience measured in years was not statistically related to the amount of uncertainty the raters experienced before, during, or after the examination rounds. Experienced raters may nevertheless be better equipped to address their own uncertainties than inexperienced raters (e.g., Ahola, 2016).

The study showed that the raters' feelings of uncertainty were not strongly related to the method of assessment (remote vs. on-site rating). However, rater uncertainty arising from the rating conditions was somewhat greater in the remote condition. The open-ended responses indicated that the raters would like to have regular on-site rating opportunities, which would allow them to concentrate on their work better and provide them with opportunities to interact with colleagues.

Our findings also show that the characteristics of examinees' test performances, such as poor handwriting, uneven performance, and performance at the borderline between two grades, create uncertainty. Prior to the beginning of the rating sessions, approximately 60 % of the raters reported that these types of performance-related factors had at least occasionally caused uncertainty in their past ratings. About half of the raters said they had also at least occasionally felt uncertain about themselves as writing raters. Additionally, a third of the raters reported uncertainty related to the potential effects of mental and physical fatigue on their ratings, and another third mentioned the effects of task interpretation and scoring criteria, and the impact of challenges related to rater training. However, the notes they made during the rating sessions showed that the uncertainty they felt when making rating decisions owed mainly to the characteristics of examinee performances. It seems that during the rating sessions, raters focus rather purely on writing performance, while outside the rating context raters reflect more on other sources of uncertainty, such as their own performance as raters (e.g., consistency, alertness). This should be further investigated. Overall, the raters rarely reported uncertainty related to rating conditions or external distractors.

It is possible that consciously attending to one's own rating in a way that differs from the usual rating process may affect rater confidence (Barkaoui, 2011). Many raters in the present study felt that participating in the research did indeed have an impact on their rating. However, the raters did not consider their participation in the study to be cumbersome, and only one rater reported that participating in the study—or monitoring their own confidence levels—was likely to have increased their feelings of uncertainty. The raters' comments indicate that the effects of participating in the study were mainly positive: for example, it added variety to the rating process, provided raters with an opportunity to reflect on their work, and gave them the chance to record and share experiences about the rating process. To ensure that raters' awareness of their own uncertainty does not cause them to over-analyze or become overly critical of their rating, future research on rater uncertainty needs to be designed carefully, and its impact on rating quality and the rating experience needs to be monitored.

It is important to understand the decision-making processes involved in testing systems that rely on human raters (e.g., Bejar, 2012). Our findings align with previous research: uncertainty is a natural element of subjective, human decision-making (Pill & Smart, 2021) and rather common in performance ratings. Uncertainty is not a continuous, overwhelming feeling but relates to specific task performances and is explained by the characteristics of the examinee performances, rating conditions, and raters themselves.

Our study also showed that uncertainty does not in itself decrease rating quality in high-stakes settings. Experiences of uncertainty can, however, be emotionally taxing, which might affect individual raters in the long run. Therefore, undue uncertainty should be addressed and reduced. Hence, information about rater uncertainty should be collected, and raters should receive support, particularly if uncertainty seems to affect the quality of their rating. Our study suggests that particularly important for rating quality and rater confidence are regular, high-quality rater training sessions (see also Davies et al., 1999; Lumley, 2005), uniform rating instructions (e.g., on the use of the criteria and their application to specific tasks), favorable rating conditions and adequate physical and mental alertness during rating. Raters should also be given regular feedback on their assessments. Receiving feedback and reflecting on one's own rating performance can have two effects on subsequent rating. Previous research shows that feedback tends to enhance rater confidence and improve rating quality but may also increase how careful raters are (e.g., Elder et al., 2005).

Advances in automated rating technologies are unlikely to remove the need for human ratings in most contexts in the foreseeable future, which is why it is important to continue investigating raters' decision-making and its underlying impetuses (see also Pill & Smart, 2021). One of the major shortcomings of the research in this area so far has to do with measurement. Most studies of uncertainty in decision making have not actually measured participants' feelings of uncertainty, and those that have, have typically relied on participants' self-reported experiences of uncertainty, which is likely to underestimate uncertainty if participants desire to portray themselves as experts who are confident in their decisions (e.g., Katz, 2002). According to Anderson et al. (2019), "other techniques for measuring uncertainty that do not rely on self-reports should be explored and validated" (p. 12). Such techniques might involve reaction-time measures, mouse-tracking, eye-tracking, and other behavioral, physiological, and affective measures that correlate with

experienced uncertainty.

The data of the study was extensive with more than 12,000 ratings and 23 writing raters. However, it should be remembered that the data were collected from only one remote and one on-site rating event and it only included intermediate level writing performances in Finnish. Therefore, it would be important to continue this line of research by conducting, for example, longitudinal studies and by expanding the research to other languages and proficiency levels. Further research could also provide more detailed information on how the uncertainty perceived by raters before, during and after assessment might differ. In addition, it could be determined whether the uncertainty is more generally connected to the important cut-off points on the rating scale or whether the rater's background might explain some of the uncertainty experienced by the rater. Scales for measuring rater certainty could be developed and validated, which may help compare the results of different studies.

Based on the present study we believe that raters' experiences of uncertainty are quite common, especially in high-stakes assessments, although they are rarely discussed. In addition to providing empirical findings, we aimed to contribute to a discussion about an assessment culture where raters can simultaneously address their possible uncertainties about assessment, gain support for reflecting on and developing their assessment skills, and maintain their face as expert raters. Even though the data used in the study come from assessment research on L2 Finnish, it is reasonable to assume that rater uncertainty manifests in a rather similar way across languages. Therefore, we believe that the results of the present study are likely to be applicable to other high-stakes writing assessment contexts. Nevertheless, it is important to recognize that Finnish, a Finno-Ugric language, differs substantially from languages such as English with respect to, for example, phonology, morphosyntax, and the processes through which speech and writing are decoded. Some of the distinctive characteristics of Finnish are its phonemic length contrasts in both the vowel and consonant systems, its relatively large number of vowels, its rich and complex nominal case system, and its nearly perfect phoneme-grapheme correspondence. It is therefore possible, though perhaps not probable, that certain sources of rater uncertainty may be related directly to the structural characteristics of Finnish or more broadly to the text types and topics used in this particular writing test.

Assessing writing in high-stakes testing is an important responsibility and raters want to do it well. This creates an interesting tension in assessment, which, as our results show, in itself can cause uncertainty in the assessment: The raters must be competent and reliable in their work, although assessment (always) involves subjectivity and emotions. Rater uncertainty is part of the assessment of writing, even if all the key features, such as rater selection and training, assessment scales, processes and instruction are of high quality. However, the features of rating quality must be monitored and maintained to ensure the validity and fairness of the assessment. This means, among other things, continuous research-based development of the testing system, the provision of rater training and discussion opportunities for the raters, continuous monitoring of rater behavior, and providing raters with feedback when necessary.

## Data availability

The data that has been used is confidential.

## APPENDIX A. The proportion of tests rated by a particular rater that included at least one task the rater was uncertain about

| The proportion of tests with uncertain assessment (s) | Raters (n) | Raters (x) |
|---|---|---|
| **less than 5%** | 8 | 36% |
| **5% - 9,9%** | 7 | 32% |
| **10% - 14,9%** | 3 | 14% |
| **15% - 19,9%** | 3 | 14% |
| **20% or more** | 1 | 5% |
| **Total** | 22 | 100% |

## APPENDIX B. Original examples from survey data in Finnish

(1) "rajatapaukset tietysti aina mietityttävät"
(2) "tekstin epätasaisuus, jolloin osa tekstistä on laadukkaampaa kuin muut osat"
(3) "huonosti luettava käsiala"
(4) "Eniten epävarmuutta aiheuttavat ohi aiheen kirjoitetut, mutta kielellisesti ansioituneet tekstit"
(5) "Joskus tuntuu, että evidenssiä on liian vähän, että kielitaidosta voisi varmasti sanoa jotain."
(6) "Hankalia ovat myös kirjoittajat, jotka ovat ehkä muilla kielillä hyviä kirjoittajia, mutta suomen taito ei vielä ole tarpeeksi korkealla tasolla."

(7) "sen pohtiminen, mitä painottaa arvioinnissa ja kuinka holistisesti osaa tehdä arvioinnin"
(8) "ns. S2-opedilemma eli ymmärtäisikö 'maallikko' tätä"
(9) "Jos edeltävästä arviointikerrasta on pitkä aika."
(10) "Joskus oman elämän asiat ja kipeä niskahartian seutu ärsyttävät ja häiritsevät arviointia, työskentelyasento on välillä tukala."

(11) “Kiireinen tunnelma. Väsymys.”
(12) “Joskus tehtävänannot eivät ole yksiselitteisiä, kuten eivät kriteeritkään.”
(13) “jos tuntuu, ettei kriteereistä ole apua (niitä voisi joskus enemmän aukaista)”
(14) “kirjoittaa runsaasti, mutta kamppailen vaikeasti luettavan käsialan takia, joka aiheuttaa lukurytmin katkeamista.”
(15) “Ajatus tulee esille, mutta vaikeaselkoisesti. ”
(16) “Sisältöä on melko paljon, mutta kielen rakenteet horjuu tosi kovasti.”
(17) “Yhteinen arviointikoulutus on ehdottoman tärkeä ja auttaa suuresti joka kerta.”
(18) “ehkä paluu lähityöskentelyyn auttaa, koska kotona arviointiympäristössä on enemmän häiriötekijöitä”

## APPENDIX C. Summary of the results of MFRM for Examinees, Raters and Tasks

| Statistics | Examinees | Raters | Tasks |
|---|---|---|---|
| Mean measure | -3.22 (1.59) | 0.00 (0.14) | 0.00 (0.06) |
| SD measure | 3.88 (0.61) | 0.95 (0.04) | 0.75 (0.02) |
| Adj. (True) SD | 3.49 (2.75)[1] | 0.94 | 0.75 |
| Min | -8.61 | -1.56 | -0.99 |
| Max | 7.99 | 1.52 | 0.74 |
| Homogeneity index | 19,695.5 (df 2480)* ** | 917.4 (df 20)* ** | 465.0 (df 3)* ** |
| Separation | 2.05 | 6.51 | 11.66 |
| Strata | 3.06 | 9.01 | 15.88 |
| Reliability | 0.81 | 0.98 | 0.99 |
| Mean Infit mean-square | 0.85 | 0.98 | 1.00 |
| SD Infit mean-square | 0.90 | 0.12 | 0.06 |
| Mean Outfit mean-square | 0.88 | 0.96 | 1.00 |
| SD Outfit mean-square | 1.15 | 0.23 | 0.18 |
| N | 2481 | 21 | 4 |

Table Notes: The rater and task facets were constrained with a mean measure of zero. Examinee results are presented with extremes, 1) without extremes. The standard errors are in parentheses. * ** p = 0.001.

## References


Ahola, S. (2022). *Rimaa hipoen selviää tilanteesta: yleisten kielitutkintojen suomen kielen arvioijien käsityksiä kielitaidon arvioinnista ja suullisesta kielitaidosta*. Jyväskylä: University of Jyväskylä. ⟨http://urn.fi/URN:ISBN:978-951-39-9005-3⟩.

Ahola, S. (2016). Puhetta arvioinnista: yleisten kielitutkintojen arvioijien käsityksiä arvioinnista. In A. Huhta, & R. Hildén (Eds.), *9. Kielitaidon arviointitutkimus 2000-luvun Suomessa* (pp. 89–109). FinLA-e: soveltavan kielitieteen tutkimuksia. Jyväskylä: AFinLA (Association Finlandaise de Linguistique Appliquée). http://journal.fi/afinla/article/view/60848

Alderson, J. C., Clapham, C. M., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press,.

Anderson, E. C., Carleton, R. N., Diefenbach, M., & Han, P. K. J. (2019). The relationship between uncertainty and affect. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02504

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51–75. https://doi.org/10.1177/0265532210376379

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Bhise, V., Rajan, S. S., Sittig, D. F., Morgan, R. O., Chaudhary, P., & Singh, H. (2018). Defining and measuring diagnostic uncertainty in medicine: A systematic review. *Journal of General Internal Medicine, 33*(1), 103–115. https://doi.org/10.1007/s11606-017-4164-1

Bogorevich, V. (2018). *Native and non-native raters of L2 speaking performance: Accent familiarity and cognitive processes. (Ph.D. dissertation)*. Department of English, Northern Arizona University.

Borders, J. C., Sevitz, J. S., Malandraki, J. B., Malandraki, G. A., & Troche, M. S. (2021). Objective and subjective clinical swallowing outcomes via telehealth: Reliability in outpatient clinical practice. *American Journal of Speech - Language Pathology, 30*(2), 598–608. https://doi.org/10.1044/2020_AJSLP-20-00234

Bosshardt, H., Packman, A., Blomgren, M., & Kretschmann, J. (2016). Measuring stuttering in preschool-aged children across different languages: An international study. *Folia Phoniatrica Et Logopaedica, 67*(5), 221–230. https://doi.org/10.1159/000440720

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Canino, E. (2001). *Posed emotional expression in brain-damaged patients across three channels of communication. (Ph.D. dissertation)*. Department of Psychology, The City University of New York.

Council of Europe. (2001). *The common european framework of reference for languages: Learning, teaching, assessment*. Strasbourg: Council of Europe,.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*, 67–96. https://doi.org/10.1111/1540-4781.00137

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press,.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch Analysis. *Language Assessment Quarterly, 2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2011). *Introduction to many-facets Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt: Peter Lang,.

Eckes, T. (2017). Rater effects: Advances in item response modeling of human ratings—Part I [Guest editorial]. *Psychological Test and Assessment Modeling, 59*(4), 443–452.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work. *Language Assessment Quarterly, 2*(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

Faraji-Rad, A., & Pham, M. T. (2013). Uncertainty increases the reliance on affect in decisions. *Journal of Consumer Research, 44*(1), 1–21. https://doi.org/10.1093/jcr/ucw073

Fitzpatrick, T., & Thwaites, P. (2020). Word association research and the L2 lexicon. *Language Teaching, 53*(3), 237–274. https://doi.org/10.1017/S0261444820000105

FNAE. (2011). *Yleisten kielitutkintojen perusteet*. Helsinki: Finnish National Agency for Education,.

Gorsuch, G., & Griffee, D. (2018). *Second language testing for student evaluation and classroom research*. Information Age Publishing Inc,.

Honko, M., Huhta, A., Neittaanmäki, R., & Jarvis, S. (2023). Instrument to study rater behaviour: Rater uncertainty and its impact on the quality of assessment in the official language examination, the National Certificates of Language Proficiency (YKI) project. *University of Jyväskylä*. https://doi.org/10.17011/jyx/dataset/88216

Hsu, H. L. (2012). *The impact of world englishes on language assessment: Perception, rating behavior and challenges. (Unpublished doctoral dissertation)*. University of Illinois at Urbana-Champaign.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvelä, T. (2014). Assessing learners' writing skills in a SLA study: Validating the rating process across tasks, scales and languages. *Language Testing, 31*(3), 307–328.

Katz, J. (2002). *The silent world of doctor and patient*. Baltimore and London: Johns Hopkins University Press,.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329–348. https://doi.org/10.1177/0265532214526177

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

Linacre, J. M. (2021). *Facets computer program for many-facet Rasch measurement, version 3.83.5*. Beaverton, Oregon: Winsteps.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang,.

Milburn, T. W., & Billings, R. S. (1976). Decision-making perspectives from psychology: Dealing with risk and uncertainty. *American Behavioral Scientist, 20*(1), 111–126. https://doi.org/10.1177/000276427602000107

Neittaanmäki, R., & Hirvelä, T. (2014). Yleisten kielitutkintojen osallistujat taustatietojen valossa. In T. Leblay, T. Lammervo, & M. (toim) Tarnanen (Eds.), *Yleiset kielitutkinnot 20 vuotta* (pp. 46–60). Helsinki: Finnish National Agency for Education.

Pill, J., & Smart, C. (2021). Raters. Behavior and training. In P. Winke, & T. Brunfaut (Eds.), *The Routledge Handbook of second language acquisition and language testing, Chapter 13.New*. York, NY: Routledge.

Posner, R. A. (2008). *How judges think*. Harvard University Press.

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and rating in oral language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty. Essays in honour of Alan Davies, 11* pp. 82–96). Studies in Language Testing.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. In *Language Testing, 25* pp. 465–493). https://doi.org/10.1177/0265532208094273

Tarnanen, M. (2002). Arvioija valokeilassa – Suomi toisena kielenä -kirjoittamisen arviointia. *Jyväskylän yliopisto: Centre for Applied Language Studies*.

Tarnanen, M. (2014). Arvioija taidon arvottajana. In T. Leblay, T. Lammervo, & M. Tarnanen (Toim.) (Eds.), *Yleiset Kielitutkinnot 20 vuotta.* (pp. 115–124). Helsinki: Finnish National Agency for Education.

Van Moere, A. (2013). Raters and ratings. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1358–1374). Wiley.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In InL. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Wind, S. A., Jones, E., & Bergin, C. (2019). Exploring patterns of principal judgments in teacher evaluation related to reported gender and years of experience. *Studies in Educational Evaluation, 61*, 150–158. https://doi.org/10.1016/j.stueduc.2019.03.011

Wind, S. A., Jones, E., & Bergin, C. (2021). Principals' severity affects teacher evaluation: statistical adjustments mitigate effects. *School Effectiveness and School Improvement, 32*(3), 413–429. https://doi.org/10.1080/09243453.2021.1892773

Youn, S.-J. (2018). Rater variability across examinees and rating criteria in paired speaking assessment. *Papers in Language Testing and Assessment, 7*(1), 32–60. ⟨https://arts.unimelb.edu.au/_data/assets/pdf_file/0006/2748255/3.-PLTA-7.1-Youn.pdf⟩.

**Mari Honko** is a Postdoctoral Researcher of Language Assessment at the Centre for Applied Language Studies, University of Jyväskylä, Finland. Her research interests focus on Finnish as a second language and later language development, language assessment, especially writing and rater behavior, multilingualism and language beliefs.

**Reeta Neittaanmäki** is a project researcher and statistician in the NCLP at the Centre for Applied Language Studies at the University of Jyväskylä, Finland. Her expertize covers quantitative research methods in the field of language testing. Her research interests focus on rater effects in performance assessment and standard setting.

**Scott Jarvis** is Professor of Applied Linguistics at Northern Arizona University. His research concentrations are second language acquisition, language assessment, and forensic linguistics. Some of his research focuses include the detection of native language influence, the use of human judges in research and assessment, and the measurement of language proficiency.

**Ari Huhta** is Professor of Language Assessment at the Centre for Applied Language Studies, University of Jyväskylä, Finland. His research interests include diagnostic L2 assessment, computer-based assessment, and self-assessment, as well as research on the development of reading, writing and vocabulary knowledge in a foreign or second language.