Language Testing in Asia

RESEARCH Open Access

Writing scale effects on raters: an exploratory study

Heejeong Jeong

Correspondence: jeongheejeong@gmail.com
Sanymyung University, 20, Hongjimun 2-gil, Jongno-gu, Seoul 03016, Korea

Abstract

In writing assessment, finding a valid, reliable, and efficient scale is critical. Appropriate scales, increase rater reliability, and can also save time and money. This exploratory study compared the effects of a binary scale and an analytic scale across teacher raters and expert raters. The purpose of the study is to find out how different scale types impact rating performance and scores. The raters in this study rated twenty short EFL essays using the two scales, completed a rater cognition questionnaire, and took part in an in-depth interview. The ratings were analyzed using a multi-faceted Rasch analysis to compare essay scores and rater statistics across scales and rater groups. The results indicated when using the binary scale, the raters spent less time and were less spread out and more consistent in their ratings. Three out of four raters replied that less mental effort was required when using the binary scale and felt more confident in their ratings. Across the two rater groups, there was a bigger shift in rating performance when using the binary scale for the teacher raters than the expert raters. This implies that scale design had a greater effect on teacher raters. The overall findings suggest that the binary scale maybe a better fit for large scale assessment with sufficient rater training.

Keywords: Binary scale, Analytic scale, Scale effect, Rater cognition, Writing assessment, Exploratory study

Introduction

With the increase in performance-based language assessment, different scales are being developed for different assessment purposes. Finding a valid, reliable, practical scale that fulfills test purposes is a challenge for test developers and administrators. Performance-based assessment, when done with human raters, is a time-consuming and expensive task; therefore, administrators want to find a scale that can reduce rating time and cost while ensuring valid and reliable results. While there has been research on the effects of tasks and raters on ratings (Schoonen, 2005), there has been limited research (Bacha, 2001; Barkaoui, 2007, 2010, 2011; O'Loughlin, 1994; Song & Caruso, 1996) on how different scales impact rating performance and scores. For this reason, I compare two different scale types (binary and analytic) that were developed on the same assessment construct (i.e., paragraph structure, content, form, and vocabulary). I investigate if there is a scale effect on the raters and how it varies across rater groups. In this study, I use a binary scale that is similar to the empirically derived, binary-choice, boundary-defined (EBB) scale originally developed by Turner and Upshur

(2002). Binary scales refer to a scale that has only two choices "yes" or "no". In the field of language testing EBB, empirically derived descriptor-based diagnostic (EDD) scales and performance decision trees are binary scale types. The advantage of binary scales is that it is easy to use and reduces working memory on the raters (Fulcher, Davidson, & Kemp, 2011). Compared to other scale types, the EBB scale has been found to be more reliable and valid (Turner & Upshur, 2002) and lessen rater cognitive load (Hirai & Koizumi, 2013). Despite the strengths of binary scales, these scales are still not widely used and not well known beyond language assessment professionals, even though it has been more than 20 years since the first publication of the EBB scale.

Being aware of the merits of the binary scale design, a binary writing scale developed by the author (Jeong, 2017a) was used for English placement test purposes at a large research university in Korea that showed success in enhancing the reliability of the writing scores. However, interviews with teacher raters reported mixed feelings regarding the scale (Jeong, 2017b). While some raters appreciated the decisive nature of the scale, others stated difficulties and unease in using a binary scale. Raters commented that the binary nature of the scale made it challenging to make decisions, and they felt uncomfortable about the way the scale forced decisions. As the developer of the binary scale, I did not expect the raters to have negative views about the scale considering its efficiency and practicality from an administrator's point of view. The different perceptions between me and the raters of the binary scale provided the motivation to pursue an exploratory study to investigate the rating performance using a binary scale compared to a traditional analytic scale across two different rater groups (expert raters and teacher raters). This study takes an exploratory research design since there is limited research that compares binary and analytic scales across different rater groups. The purpose of the study is not to generalize the findings but rather to gain preliminary insights to a question that arose from my experience working as a test administrator.

While there have been a few studies (e.g., Hirai and Koizumi, 2013; Kubota, 2018) that have compared rater performance using a binary scale and an analytic scale, I do not know of any studies that have used language assessment professionals rather than experienced teachers as the expert rater group for comparison.

Literature review

Different rater background effects

In scale comparison studies, raters who are the users of the scales are important factors that affect rating results. Raters depending on their background (e.g, native language, rating experience) can be divided into different groups. Among studies concerning rater background, the issue of native and non-native speakers (NNS) and their rating patterns have been explored by multiple scholars. Some studies found significant differences based on the L1 background (Zhang and Elder 2014; Gui, 2012) while others found no significant differences (Johnson and Lim's 2009; Barkaoui, 2011). Research that found a minimal effect of L1 language background had NNS raters who possessed near-native language fluency (Johnson and Lim, 2009; Barkaoui, 2011). Johnson and Lim's (2009) study looked at the rater language background effect on the rating of performance assessment in the MELAB writing test. The study compared 4 NNS raters (Spanish speaker, Tagalog speaker, Chinese speaker, Korean speaker) who had native or native-like proficiency with 15 native raters. The findings of this study report no patterns of language-related bias in the ratings. The authors state, "rater language background [effect] can be minimized and made a non-factor in the scoring of writing performance assessments (p.502)". They also state both native non-native raters can be educated to use a rubric that makes it difficult to distinguish rating behavior across groups.

Another rater background factor that plays an important role in rating patterns is rater experience. Similar to L1 background, findings by raters with different levels of rating experience are mixed. Royal-Dawson and Baird (2009) found no significant differences while Sakyi (2003) found experience raters to score faster and consider a wider variety of language features less experienced raters. In Barkaoui's (2011) study, experienced raters presented a higher exact rater agreement (26%) compared to the novice raters (20%) and were also stricter in the ratings. The acceptable fit was higher (62%), and the student separation index was higher for the experienced raters. Thus, the experienced raters were more homogeneous as a group. On the other hand, novice raters were more lenient and exhibited more misfit. There were more intra-rater variabilities in their ratings, and more rater variations were detected for novice raters. Barkaoui's study concludes that rating experience did play an important role in rating patterns and rating scores.

Binary scales

In addition to rater background, another key factor that influences student essay scores and rater performance are scale types. Among different scale types (e.g., holistic, analytic, primary, multiple trait), the binary scale is a type of nominal scale that consists of two possible values. Binary scales are developed using the performance-data driven approach; therefore, it is constructed based on empirical student data. In the field of language testing, EBB, empirically derived descriptor-based diagnostic (EDD) checklist, and performance decision trees are binary types of scales. In all binary scales, a series of binary questions (yes/no) are given to the rater to make judgments. Compared to other scales, binary scale types are found to be simple, easy to use, and do not place a heavy burden on the rater's working memory during the rating process (Fulcher, Davidson, & Kemp, 2011).

The EBB scale which is one of the oldest and widely used binary scale was originally developed by Upshur and Turner (1995), has been reported to be valid and reliable, especially for speaking assessment (Turner & Upshur, 2002). Another type of binary scale is the empirically derived descriptor-based diagnostic (EDD) checklist developed by Kim (2010). Kim's EDD checklist consists of 35 yes/no questions addressing various ESL writing skills such as content fulfillment, organizational effectiveness, grammatical knowledge, vocabulary use, and mechanics. Raters in Kim's study appreciated the practicality of the scale but also expressed concern regarding the lack of a continuum. They commented that "yes" in one aspect of writing does not imply mastery of that writing skill. They were also concerned about the increased cognitive load in the writing process. Raters felt dichotomizing writing competence into "yes" or "no" choices gave a heavy weight to their cognitive load. Unlike traditional scale types such as holistic or analytic, binary scales can be considered a new type of scale in the field of language testing and research on the effects of binary scales on raters is lacking.

Scale effects on raters

Previous scale comparison studies have focused heavily on holistic or analytic scale types (Bacha, 2001; Barkaoui, 2010, 2011; O'Loughlin, 1994; Song & Caruso, 1996), and only recently there have been studies that have compared other types of scales such as binary scales. Hirai and Koizumi's (2013) study researched the reliability, validity, and practicality of three rating scales. Although the focus of the study was on the validation of two EBB scales (i.e., EBB1 and EBB2) developed for rating a speaking test, the study also compared rater performance when using the EBB2 scale and a multiple-trait (MT) scale. The multiple-trait scale, which resembles the design of an analytic scale, measured the same assessment construct as the EBB2 scale with the only difference being the scale design. According to the findings, the MT scale showed a lower exact rater agreement ratio compared to the EBB2 scale. Out of the seven raters, two misfit raters were identified. This implies raters had more difficulty rating consistently using the MT scale. The authors state that raters had difficulty rating consistently using the MT scale. Hirai and Koizumi assumed that the design of the MT scale might have created too much of a cognitive demand on the raters by showing all score descriptors at once, which may have led to fluctuating ratings across the five levels. On the other hand, the EBB2 scale demonstrated the highest person separation ratio and the greatest discrimination power. The statistics concerning scale reliability were better for EBB2 than the MT scale. However, when the authors investigated the practicality of the two scales, four out of the seven raters reported spending less time on scoring performance using the MT scale, and five out of the seven raters found it easier to use. Overall, despite the higher rater reliability, the raters in this study found EBB2 scales more time consuming and difficult to use. The authors write that the format of the MT scale could have made it easier for raters to use the scale, and the unfamiliarity of the EBB2 scales could have required more rating time. In terms of practicality, the authors believe the MT scale to be the best choice. They conclude that the EBB2 scale could have required raters to be more cautious about their ratings.

Recently, Kuobota (2018) also conducted a rating scale comparison study using an EBB scale. In this study, Kubota compared the impact of the ESL Composition Profile (Jacobs et al., 1981) scale and an EBB scale developed for the study. The purpose of the study was to explore the effects of rating scales on experienced and inexperienced raters. The study was conducted using a mixed-methods approach. The authors used MFRM to analyze the essay data rated by three experienced and two inexperienced raters. Results showed that the inexperienced raters assessed the essays more leniently than the experienced raters, and grammar was rated most severely. In terms of scale perception, interview findings showed that even though the EBB scale design was new to the inexperienced raters, they found it easy to use. These raters appreciated the hierarchical set of descriptors of the scale. Kubota concludes that untrained raters can benefit from the explicit and hierarchical design of the EBB scale.

As found in previous EBB scale comparison studies, rater perceptions on this binary type of scale were not consistent (Hirai and Koizumi, 2013; Kuobota, 2018). Some raters (Hirai and Koizumi, 2013) said it was more time consuming and difficult to use, while others thought it was easy to use, even with inexperienced raters (Kuobota, 2018). This gap in the literature calls for further research on rating performance when using binary and analytic scale types. This study also examines the effects of two different scales (binary scale vs. analytic scale) with two different rater groups (teacher raters vs. expert raters) and their interactions on estimates of student writing scores, rater agreement, rater severity, and self-consistency.

Binary scales have been found to be reliable and valid but there have not been many studies related to this scale type. Previous scale effect studies were limited to holistic vs. analytic scales. Recently, rater effect of different scales such as binary scales have been conducted but the number of research is still scarce and the findings are not consistent. This study is a further development of previous work (Barkaoui, 2010, 2011; Hirai & Koizumi, 2013; Kuobota, 2018) that compared different scale types. Through this study, I hope I can find the strengths and weaknesses of binary scales in contrast to other scales and its effect across different raters groups.

Research Questions

(1) What are the effects of binary and analytic scales?

(2) Is the scale effect different for expert raters vs. teacher raters?

Method

Participants

Four raters (Table 1), two expert raters (Young and Sue), and two teacher raters (Fred and Sean), took part in the study. The expert raters were non-native professors in the field of language assessment who possessed near-native language abilities. Young and Sue both took courses related to language assessment, educational measurement, and statistics courses as part of their doctoral study. This rater group had extensive experience, not only in the rating process but also in developing and validating rating scales. A detailed description of the participants are stated in Table 1. The teacher–rater groups (Table 1) were university instructors who were currently teaching general English courses in an EFL context. They were native speakers (NS) of English and had a master's degree in areas related to EFL/ESL teaching. This group of teachers will likely be the users of the scales. To find out how teacher raters perceived scales differently in

Table 1 Rater profiles

	Teacher rater		Expert rater	
	Fred	Sean	Sue	Young
Assessment experience	Placement test essay rating, classroom based	Classroom based	English placement test developer, administrator	Nationwide performance assessment designer, rater trainer
Rater training experience	No	No	Yes	Yes
Language assessment course completion	No	No	Yes	Yes
Post-graduate study related to language assessment	No	No	Yes	Yes

contrast to expert raters, it was important to have a clear distinction between the two rater groups.

The number of raters in this study was small and they also possessed different L1 s. In my research context, recruiting NS instructors for the teacher rater group was not very difficult, but it was challenging to recruit LT professionals for the expert rater group and almost impossible to find NS LT professionals. This resulted in small sample size and different language backgrounds. The teacher raters were native speakers and the expert raters were not. This is a concern but it reflects the character of the educational context I am part of. Small sample size and different L1 backgrounds are the concerns and limitations to the study but as stated earlier, the purpose of this exploratory study is not to generalize the findings but to explore different scale perceptions across different rater groups. Even though data was collected from a small number of participants, I tried to incorporate various methods (essay rating, survey, interview) to collect in-depth information.

Binary scale and analytic scale

Two different scales were used for this study: A binary scale (Appendix 2) that was developed and validated by the author (Jeong, 2017a) and an analytic scale (Appendix 1) that was developed on the same construct as the binary scale. The two scales were designed to represent the same writing assessment construct and measure the same writing traits. The only difference was the layout of the scale. The analytic scale was reviewed by a language testing expert to ensure that the assessment constructs were similar. This expert did not participate in the main study.

Data collection process and analysis

Raters assessed twenty student essays written for an on-line placement test. The essays were produced under real exam situations and were written within 30 min. The student writings consisted of 30% (n = 6) low level, 40% (n = 7) mid level, and 30% (n = 7) high level essays. The essays used in this study were from a larger data set and were previously rated. The reason for choosing three levels was to have a sample of essays that represented a wide range of writing proficiencies. The rating process for the two groups was the same. The two rater groups rated twenty short EFL writing samples (100 to 200 words) using both scales. The twenty essays were divided into two sets (set 1, set 2) each consisting of 10 different essays from various proficiency levels. To ensure a counter-balance of ratings, in the first rating session (rating 1) raters assessed set 1 with the binary scale and set 2 with the analytic scale (Table 2). For the second rating session (rating 2), raters rated set 2 with the binary scale and set 1 with the analytic scale. The essays were divided into two sets, to make sure the raters rated each set with both scales and also in reverse order. For more details, please refer to the rating guideline in Appendix 3.

Table 2 Rating process

Rating 1	Rating 2
Set 1- Binary scale	Set 2- Binary scale
Set 2- Analytic scale	Set 1- Analytic scale

No rater training was given for the study. In addition to rating the essays, the raters measured the total time spent when rating the essays for the two different scales. The time was self-measured for each set per rating session (e.g., rating 1- set 1- 29 min). The raters were advised to set a time difference between the two ratings to avoid a memory effect.

The essay scores were analyzed using the FACETS 3.71 (Linacre, 2014) program to check psychometric measures related to rater reliability, rater variability, and test taker discrimination. The program provides estimates for each facet (e.g., student, rater, scale type) and displays it on a single logit scale (McNamara, 1996). The severity and leniency in rating patterns can be identified through the FACETS program. Raters in this study rated the same essays; therefore, it was a fully crossed design. A four facet (student, rater, rater group, criteria) model was used for the analysis. Using the FACETS program, three MFRM models were examined: scores analyzed with the binary scale, analytic scale, and with both scales.

After completing both rating sessions, the raters immediately filled out a questionnaire (Appendix 4) that measured the cognitive effort required in doing the ratings. The questions were based on the self-reporting cognitive scale developed by Paas, Ayres, and Pachman [2008]. After finishing Phase 1 (quantitative data) of the study, all the data were analyzed by the researcher. The interview questions were developed on the findings of the FACETS analysis and questionnaire results. Using these findings, individual rater interviews were then conducted (phase 2) that asked the rater's overall rating experience using the two scales. All the interviews were digitally recorded and analyzed based on the similarities and differences in the scale effect across the rater groups.

Results

RQ. (1) What are the effects of binary and analytic scales?

Student statistics

To find out if the two rating scales assessed the same writing construct, a Wilcoxon signed-rank test was conducted using SPSS 23. The test detected no significant differences in the ranking of the student's writing ability across the binary and analytic scales ($Z = -1.084$, $p = 0.278$). This means that the two scales are measuring the same assessment construct. However, when the Wilcoxon signed-rank test was done for the individual raters, the results were not the same. For the first set, Young ($Z = 0.00$, $p = 1.00$), Sue, ($Z = 0.69$, $p = 0.490$), and Sean ($Z = -1.382$, $p = 0.167$) showed no significant differences, which implies that the rating was similar across the two scales, but for Fred ($z = -2.620$, $p = 0.009$), the Wilcoxon signed-rank test did show a significant difference between the scores from the binary scale compared to the analytic scale. This means that Fred interpreted the scales differently. For the second essay set (set 2- analytic scale first, binary scale later), no raters showed a significant difference across the two scales, which indicates an improvement in using the scales as a result of practice.

According to FACETS, the overall essay scores were not statistically significantly different for the two scales. Student essays, when rated by the binary scale, showed a slightly higher fair average (2.63) than the analytic scale (2.58), but this was not significant (fixed chi-square = 1.3, $p = 0.26$). The student separation index was 2.44 for the analytic scale and 2.52 for the binary scale. In terms of the infit statistics, the overfit and misfit mean squares were all 5% higher for the analytic scale, resulting in a higher acceptable fit (10%) for the binary scale (Table 3).

Scale statistics

The specific assessment criteria (paragraph structure, content, form, and vocabulary) showed similar patterns in severity for both rating scales. Form was the most difficult criterion for both the binary (0.99 logits) and the analytic (0.93 logits) scale followed by paragraph structure but for vocabulary and content, the order was the reverse. Content was the easiest criterion for the analytic scale (− 0.71) whereas vocabulary (− 0.43) was the easiest for the binary scale. The separation index was higher for the binary scale (3.01) than the analytic scale (2.68). The two scales (Appendixes 1 and 2) that were used in this study covered a 4-point range (1–4). Students were given 1 point for the lowest performance for each category and 4 for the highest. One noticeable difference between the two scales was the central tendency of categories 2 and 3 for the analytic scale. When using the analytic scale, raters assigned 87% (278 counts) of their scores either a 2 or 3, but for binary, this was reduced to 80% (256 counts). This shows that more high-end and low-end scores were given for the binary scale (20%, 64 counts) in contrast to 13% (42 counts) for the analytic scale.

Rater statistics

Raters took less time rating (Table 4) with the binary scale (M = 27.37 min, SD = 4.27 min) compared to the analytic scale rating (M = 30.12 min, SD = 6.86 min). The binary scale showed a separation index of 2.14. This means that raters can be separated into two different levels. In contrast, the analytic scale showed a separation index of 4.21. When assessing with the analytic scale, raters can be separated into four different levels. From these findings, we can say that the raters were more widely distributed when using the analytic scale compared to the binary scale. The exact rater agreement was higher for the analytic scale (44%) than the binary scale (37.3%). In terms of the fit statistics, all the raters were within the acceptable fit range for the binary scale, but there was one misfit and one overfit for the analytic scale. Except for the exact rater agreement, the rater statistics were better for the binary scale compared to the analytic scale. Raters spent less time and were less spread out and more consistent in their rating when using the binary scale (Table 4).

Table 3 Student statistics

	Binary	Analytic
Fair average	2.63	2.58
Student separation index	2.52	2.44
Overfit	15% (*n* = 3)	20% (*n* = 4)
Acceptable	65% (*n* = 13)	55% (*n* = 11)
Misfit	20% (*n* = 4)	25% (*n* = 5)

Table 4 Binary and analytic scale comparison

	Binary	Analytic
Time	27.37 min	30.12 min
Rater exact agreement	37.3%	44%
Rater separation index	2.14	4.21
Rater fit	All within the acceptable fit (0.82–1.15)	1 misfit (0.62), 1 overfit (1.36)

RQ. (2) Is the scale effect different for expert raters vs. teacher raters?

Rating scale statistics between the rater groups

To investigate the scale effect for the two rater groups, the binary and analytic scales were compared for both the teacher rater and expert rater with regard to time, fair average, exact rater agreement, and rater separation index (Table 5). There was a 7.25-min difference in the rating time for the binary scale and a 6.75-min difference for the analytic scale across the rater groups. While both rater groups spent less time when rating with the binary scale, the time difference was greater for the binary scale. For the fair average, the difference was 0.14 for the binary scale, but 0.09 for the analytic scale across rater groups. When rating with the binary scale, expert raters showed a 16.2% more exact agreement rate than the teacher raters in contrast to 7.5% more exact agreement rate for the analytic scale. The separation index was wider for the analytic scale across the raters. In comparing rater statistics across the rater groups for the two scale types, there were greater differences across groups when rating with the binary scale compared to the analytic scale. More differences in time, student fair average scores, and exact rater agreement were found. The only criterion that showed a greater difference for the analytic scale was the rater separation index. This implies that teacher raters experienced more challenges in rating with the binary scale compared to the analytic scale.

Rater cognition questionnaire

After the ratings were completed, the raters immediately filled out a short rater cognition questionnaire that investigated the degree of rater cognitive load (Appendix 4) required when doing the ratings. The findings showed that three out of the four raters felt that more mental effort was required when using the analytic scale. In terms of ease, three raters (Frank, Young, and Sue) said that the binary scale was easier to use, but Sean felt it was the same for both scales. For rating confidence, three raters felt more confident when rating with the binary scale. Overall, the questionnaire results indicated that all raters except for Sean perceived that the binary scale was easier to use,

Table 5 Scale comparison across rater groups

	Binary			Analytic		
	Teacher rater	Expert rater	Differences	Teacher rater	Expert rater	Differences
Time	*M* = 31, SD = 2.91	*M* = 23.75, SD = 1.29	7.25	*M* = 33.5, SD = 7.43	*M* = 26.75, SD = 4.02	6.75
Fair average	2.72	2.58	0.14	2.63	2.54	0.09
Exact agreement	30%	46.2%	16.2%	33.7%	41.2%	7.5%
Rater separation index	2.37	1.02	1.35	5.91	2.07	3.84

increased rating confidence, and lowered the cognitive load raters experienced during the rating process.

Rater interview group findings: similarities

Once the ratings and questionnaires were completed, rater interviews were conducted either face to face or over the phone that lasted 30 to 40 min. The interview data were analyzed based on the similarities and differences between the two rater groups.

The findings showed that it was the first time to use a binary scale for the raters. The expert raters were aware of this scale type but did not have experience using the scale. For the teacher raters, this was their first time to encounter a binary scale. Unlike the binary scale, the analytic scale was familiar to all raters, and everyone had experience using it. The raters in this study commented that the binary scale required more time to understand at the beginning but with more practice, they became familiar with using the scale. A shared feature everyone agreed on concerning the two scales was when it should be used. Although the raters found the binary scale easy to use and efficient, they preferred to use this scale for large scale purposes such as placement testing whereas the analytic scale was more appropriate for classroom assessment purposes as it helped to identify the weak areas of students. They indicated two reasons for this. First was the flexibility of the scale. Unlike the binary scale, which was fixed in terms of scale structure, raters found that the analytic scale allowed more flexibility in the rating process. They felt this element was important in classroom-based assessment. Sue stated that if she used the analytic scale, she could easily control the number of A's and B's when grading on a curve. Another reason why the analytic scale fit better for classroom assessment purposes was the feedback factor. Raters said that in a classroom context, the analytic scale would be more appropriate when giving feedback to the students. From the rater interviews, we can see that, regardless of group differences, the scale choice was dependent on the assessment purposes.

For mental effort, three raters (Sue, Young, and Fred) replied that the binary scale required less mental effort, giving them more confidence in their ratings. These raters felt there was less confusion in their ratings, so they found it easier to make a decision. When rating with the analytic scale, the raters responded they had to think of multiple areas and experienced more hesitation and rating conflict. Raters also felt the urge to assign a mid-point. Unlike the three raters, Sean shared a different opinion about the binary scale. Sean said he found himself double checking more using the binary scale because the scale was so simple and clear that it was not easy to be confident in his ratings. This resulted in Sean having a lack of trust in his rating decisions making him feel uncomfortable and more nervous. Having no middle ground lowered his confidence level in his ratings. Sean said he was more confident with his analytic ratings since he felt the ratings could be more precise. For other raters, they liked the decisive nature of the binary scale and having no "maybes" raised their rating confidence.

Rater interview group: differences

While the experiences of the binary and analytic scale were similar across both rater groups, there was a difference in how they understood the scale. An area teacher raters and expert raters showed differences was the construct of the scale. The purpose of the study and the information of the assessment constructs of the scales were not given to the raters. Yet, the expert raters in this study knew that the two scales measured similar assessment constructs, but the teacher raters were not aware of this and thought the analytic scale included more assessment criteria. This means that the layout and design had an effect on the teacher raters. This rater group was more sensitive and influenced by scale design. Sean said, "I felt there was more content in the analytic scale. When I was looking at the [binary scale], I was thinking, a lot of things that I look for are not here [in the binary scale]." Sean continued to say, "I think it [analytic scale] has more verbiage, more to look at, more guidance, than the binary. I felt, if I were to analyze the essays more carefully, I would be doing more precise evaluation using the analytic scale." Another difference between the rater groups was the decision making process. The expert raters welcomed the simple "yes" and "no" structure, but Sean felt uneasy about this simple form. He felt it was too simple and lacked confidence in his ratings. Fred also had trouble at the beginning when using the binary scale. Fred's challenge in using the scale was evidenced by the significant difference in the Wilcoxon signed-rank test when rating the first set of essays. Fred described his first experience using the binary scale as follows, "I was at first a little bit confused. In the beginning, but by the third and fourth time I got used to it. … In the beginning, it was foreign, once I understood it, it became easy. First I had to understand the scale because I have never used the binary scale." The interviews and rater statistics clearly showed that the teacher raters experienced a bigger learning curve for the binary scale than the expert raters.

Summary and discussion

Scale effects

The Wilcoxon signed-rank test and FACETS findings showed the two scales measured similar assessment constructs. The Wilcoxon test detected no significant differences in the ranking of the student's writing ability across the binary and analytic scales, and student scores student separation index was also similar (Table 3). This finding was expected since the two scales were developed on the same criteria and the only difference was the scale design.

The same student rankings and similar student scores did not mean the two scales had the same effect on raters. First, raters spent less rating time when using the binary scale (M = 27.37 min) than the analytic scale (M = 30.12 min). In a rating context, rating time is an important factor and is directly related to the efficacy of the scale. If the same ratings can be done in a shorter time, with better or minimal difference in rating quality, test administrators will more likely choose a binary scale. Next, the rater separation index was smaller (2.14) when using the binary scale in contrast to the analytic scale (4.21). Also while all raters were within the acceptable fit range for the binary scale, there was 1 misfit and 1 overfit for the analytic scale. Third, the overall exact rater agreement was higher for the binary scale (46.2%) compared to the analytic scale (41.2%). These findings are supported by Hirai and Koizumi (2013) who also found a lower exact rater agreement and more misfits for the analytic MT scale in contrast to the binary EBB scale.

As summarized in Table 3, better rater statistics were evidenced in most areas when rating with the binary scale. An explanation for these findings could be that when using the binary scale, raters only need to choose between a "yes" and "no." This could have lowered the mental effort required in the rating process and produced more reliable ratings. Similar to Hirai and Koizumi (2013)'s results, the simple structure of the binary scale could have lowered the working memory load of the raters. This explanation was supported by the rater cognition questionnaire findings and rater interviews. All raters except Sean reported less mental effort was needed when rating with the binary scale and stated the binary scale was easier to use.

Rater group effects

Overall group differences did exist for teacher raters and expert raters (Table 5). Similar to findings from previous studies, expert raters were able to do the ratings faster (Sakyi, 2003) and showed more rater exact agreements and higher acceptable fit (Barkaoui, 2010). Teacher raters spent more time, assigned higher student scores, and reported less exact rater agreement when rating with the binary scale than the expert raters (Table 5). However, the difference between the two groups was bigger for the binary scale. First, there were more group differences (Table 5) for rating time (binary 7.25 min, analytic 6.75 min), student score fair average (binary 0.14, analytic, 0.09), and exact rater agreement (binary 16.2%, analytic 7.5%) when using the binary scale than the analytic scale. The reasons for theses FACETS results may be due to the foreign design of the binary scale. As reported in the rater interviews, the binary scale was new to the teacher raters and they had trouble using the scale especially in their first ratings. Sean said he was uncomfortable using the binary scale and felt that there should be more content written in the scale in order for him to feel more comfortable with his ratings. Frank said it was challenging to use the binary scale when he first saw it. On the other hand similar to my expectations, the two expert raters who were aware of the binary scale design welcomed the binary scale from the beginning. Young said she really liked the binary scale because it can reduce external factors that can influence ratings.

One reason why non-assessment professionals like Sean may feel uncomfortable with this scale could be the very simple design of the scale. There could be a difference in the confidence level of a rater depending on their level of expertise. Raters who are more experienced and are able to justify their ratings based on their educational background and expertise may not need a detailed scale to explain their ratings. However, raters like Sean who are less experienced might need more evidence to rely on to support their ratings. Other reasons for teacher raters showing difficulty or expressing discomfort in using the binary scale could be that they were less aware of the two different worlds of assessment: large scale and small scale. While binary scales have been developed for classroom assessment purposes, this scale type has been more widely used for large scale assessment purposes (Author, 2017). Unlike the expert raters, Fred and Sean had a limited view of the assessment context. They referred only to the small scale classroom situation when describing the application of the two scales. The implication is that teacher raters who do not have a wide range of assessment experiences may have difficulty understanding the purpose and use of different scale types.

Overall scale design had a bigger effect on the teacher raters than the expert raters. The rater separation index was bigger (binary 2.37, analytic 5.91) for the teacher raters than the expert raters (binary 1.02, analytic 2.07). This result is similar to Barkaoui's (2010) study that reports more scale type influence for novice raters compared to experienced raters. Unlike expert raters, who have the capability to look beyond the stylistic features of a scale, teacher raters were directly impacted by it. Sean commented that there was more content in the analytic scale. During the interview, I pointed out to Sean that the two scales were actually developed on the same assessment constructs. After hearing this he commented, "It's purely cosmetic. It looks like there is more meat in it. The way it is written and structured, you feel like you are looking for more."

Regardless of the different scale design, the expert raters knew there was little difference in the assessment constructs across the scales. When developing scales, attention is usually given to the assessment criteria, descriptors, and scoring methods but not to the layout. From this study, we can say that in addition to the traditional factors considered in scale development, the design factor also has an impact on raters.

Characteristics of a good scale

There is general agreement on the shared characteristics that constitute good scales. First, these scales report better rater statistics (e.g., rater reliability, rater separation ratio, exact agreement ratio) and student statistics (Barkaoui, 2011; Hirai & Koizumi, 2013; Knoch, 2009). Next, such scales impose less cognitive demand on the teacher rater (Bakaoui, 2010; Hirai & Koizumi, 2013) and require less rating time. Through the findings of this study, we can see that the binary scale includes the features of a stronger scale. It shows better student discrimination (student separation index binary 2.52, analytic 2.44), less rater variability (rater separation index binary 2.14, analytic 4.21), better rater fit (binary no misfit, analytic 1 overfit, 1 misfit), and less rating time (binary 27.37 min, analytic 30.12 min). This means, when deciding on a scale type, the binary scale may be a better choice for both test administrators and raters. Next, from the rater cognition questionnaire and rater interview, three raters said that the binary scale required less mental effort in comparison to the analytic scale. Findings from various sources in this study suggest that the binary scale is a stronger scale than the analytic scale.

Promoting the binary scale

From this study, I found that the binary scale is still not well-known beyond the language assessment community and not widely used. None of the raters who took part in the study had experience using the scale, and the teacher raters did not know of its existence. However, when given the opportunity to use such a scale type, the raters did appreciate the efficiency of the scale and shared a common belief in it as an effective scale, particularly for large scale assessment purposes. From this study, to use binary scales several conditions should be met. First is the need for sufficient rater training. The teacher raters in this study did not have experience in large scale assessment but did have more than 10 years of experience as language teachers. In other studies (e.g., Barkaoui 2010, 2011), these teacher raters are often labeled as experienced raters in contrast to novice raters. Considering the difficulties, the two teacher raters had when first using the scale, more rater training is needed when using this scale type with first time users. Another challenge that limits the use of the binary scale is the difficulty of developing one. Unlike generic scales that can be applied to different tests, the binary scale is developed upon empirical data. This increases the validity of the scale but also limits the generalizability. Different scales need to be developed for different assessment purposes. Developing a scale is definitely time-consuming and requires a lot of effort and expertise. However, once a scale is developed as shown in this study, rating time can be reduced, which can compensate for the additional time needed in developing a scale. Test administrators should be aware that the advantages of a binary scale can outweigh the disadvantages in the development stage.

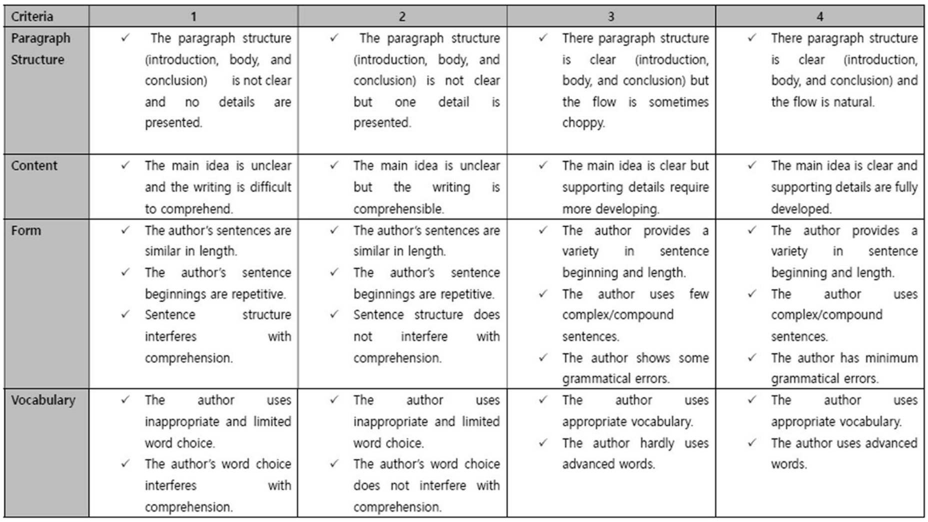
Conclusion and implications for future studies

As stated before, this is an exploratory study and data was collected from a small number of participants. Even though the expert rater and teacher rater group were selected based on their educational and teaching backgrounds, the number of raters was small ($N = 4$). This is a limitation in the study because the findings could represent individual rating patterns and perceptions rather than overall group differences. I do advise readers to interpret the findings referring to different groups with more caution. For future studies, I recommend recruiting a larger group of raters and breaking down the raters into three groups: expert, experienced, and novice. Rater groups are often divided into two groups (e.g., experienced vs. novice) but from this study, we can see that the expert rater group had different perceptions than the non-language assessment professionals. In reality, expert raters will probably not be recruited as raters, but I believe it is important for experts in language assessment to gain perspective by participating in a study with lay raters to see how rating scales are used in the rating process and how it can be different from the developer's perspective.

Another area that should be looked into more deeply in future studies are issues related to rater cognition. I lightly touched on this topic through the rater cognition questionnaire and rater interviews, but there is a limit in the data that can be collected through these methods. To collect more in-depth data on this topic, I propose to investigate rater cognition through think-aloud protocols. Think aloud studies (e.g., Lumley, 2002; Wolfe et al., 1998) have been used to research the rating process and could generate more meaningful results that could allow us to better understand the cognitive activities going on through the rating process.

Despite these limitations, I believe the results of findings of the study gave preliminary insights in binary scale studies. Test administrators are always in search of a scale that is more efficient and reliable to use. Training and using human raters is expensive. If there is a scale that can cut down on rating time and still maintain rater reliability, it would be beneficial to use it. A binary type of scale could be a solution to these problems. I believe the study findings contributed in providing more in-depth knowledge on how different rater groups use and perceive binary scales. I began this exploratory study to find answers why such a good scale (a.k.a., binary scale) from my perspective was not widely used in language assessment. Through this study, I did find possible answers to my question and hope to see more future research in this area.

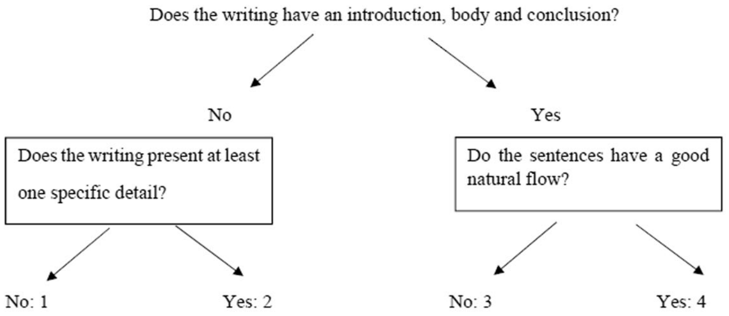
Appendix 1



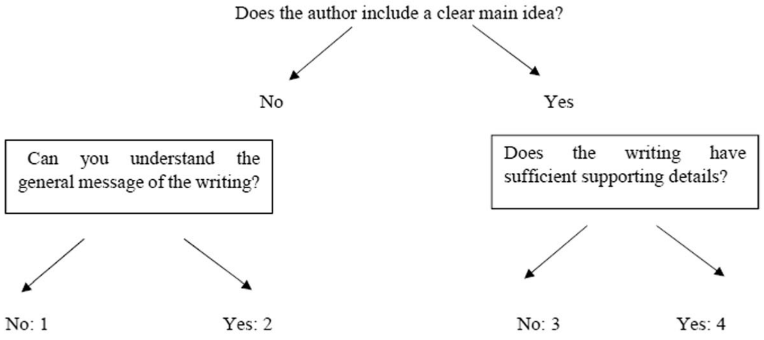
Criteria	1	2	3	4
Paragraph Structure	✓ The paragraph structure (introduction, body, and conclusion) is not clear and no details are presented.	✓ The paragraph structure (introduction, body, and conclusion) is not clear but one detail is presented.	✓ There paragraph structure is clear (introduction, body, and conclusion) but the flow is sometimes choppy.	✓ There paragraph structure is clear (introduction, body, and conclusion) and the flow is natural.
Content	✓ The main idea is unclear and the writing is difficult to comprehend.	✓ The main idea is unclear but the writing is comprehensible.	✓ The main idea is clear but supporting details require more developing.	✓ The main idea is clear and supporting details are fully developed.
Form	✓ The author's sentences are similar in length. ✓ The author's sentence beginnings are repetitive. ✓ Sentence structure interferes with comprehension.	✓ The author's sentences are similar in length. ✓ The author's sentence beginnings are repetitive. ✓ Sentence structure does not interfere with comprehension.	✓ The author provides a variety in sentence beginning and length. ✓ The author uses few complex/compound sentences. ✓ The author shows some grammatical errors.	✓ The author provides a variety in sentence beginning and length. ✓ The author uses complex/compound sentences. ✓ The author has minimum grammatical errors.
Vocabulary	✓ The author uses inappropriate and limited word choice. ✓ The author's word choice interferes with comprehension.	✓ The author uses inappropriate and limited word choice. ✓ The author's word choice does not interfere with comprehension.	✓ The author uses appropriate vocabulary. ✓ The author hardly uses advanced words.	✓ The author uses appropriate vocabulary. ✓ The author uses advanced words.

Appendix 2

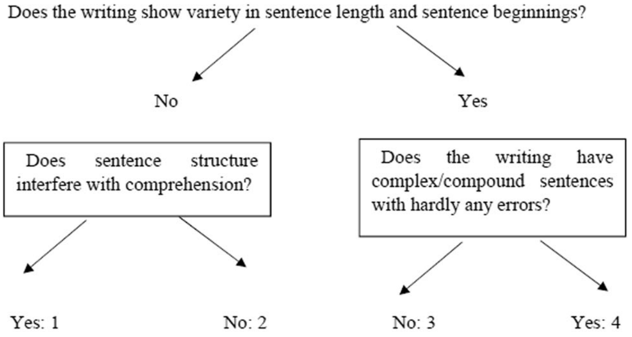
1. **Paragraph Structure**


Does the writing have an introduction, body and conclusion?
No
Yes
Does the writing present at least one specific detail?
Do the sentences have a good natural flow?
No: 1
Yes: 2
No: 3
Yes: 4

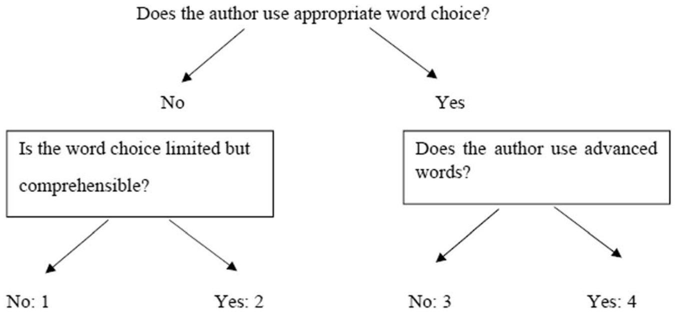
2. **Content**


Does the author include a clear main idea?
No
Yes
Can you understand the general message of the writing?
Does the writing have sufficient supporting details?
No: 1
Yes: 2
No: 3
Yes: 4

3. **Form**


Does the writing show variety in sentence length and sentence beginnings?
No
Yes
Does sentence structure interfere with comprehension?
Does the writing have complex/compound sentences with hardly any errors?
Yes: 1
No: 2
No: 3
Yes: 4

4. **Vocabulary**


Does the author use appropriate word choice?
No
Yes
Is the word choice limited but comprehensible?
Does the author use advanced words?
No: 1
Yes: 2
No: 3
Yes: 4

Appendix 3

Rating guidelines

1. Before rating, please familiarize yourself with the two scales (analytic, binary).

2. The student essays in this study are from an on-line English placement test for college freshmen. The essays consist of students from different writing proficiency levels.

The following is the writing prompt.

Please write an essay for the following topic.

Nowadays media pays too much attention to the personal lives of famous people such as celebrities. Do you agree or disagree with this statement? Present specific reasons to support your answer.

3. Rate the first student essay (Set 1) set using the binary scale. Please input your ratings in the Excel file titled ‘1st_Binary Set 1’.

Please measure the total rating time for each scale. (e.g., Set 1- Binary- 45 minutes)

4. Rate the second student essay set (Set 2) using the analytic scale. Please input your ratings in the Excel file titled ‘1st_Analytic Set 2’.

Please measure the total rating time for each scale. (Set 2- Analytic- 30 minutes)

Set 2 Analytic Total time:

5. After a few days later, please rate the student essays one more time using different scales. Please input your ratings in the Excel file titled ‘2nd_Binary Set 2, 2nd _Analytic Set 1’.

(There should be some time difference between the two ratings to avoid memory effect.)

Set 2 Binary Total time:

Set 1 Analytic Total time:

Appendix 4

Rater cognition questionnaire

1. In rating the essays with the analytic scale, I invested

a. Very low mental effort, b. low mental effort, c. neither low nor high mental effort

d. high mental effort e. very high mental effort

2. In rating the essays with the binary scale, I invested

a. Very low mental effort, b. low mental effort, c. neither low nor high mental effort

d. high mental effort e. very high mental effort

3. When rating the student essays with the analytic scale, I felt _______ about my ratings.

a. not at all confident, b. slightly confident, c. somewhat confident, d. moderately confident, e. extremely confident

4. When rating the student essays with the binary scale, I felt _______ about my ratings.

a. not at all confident, b. slightly confident, c. somewhat confident, d. moderately confident, e. extremely confident

5. Which scale was easier to use when rating the essays?

a. analytic, b. binary, c. same

Abbreviations

EBB: Empirically derived, binary-choice, boundary-defined; EDD: Empirically derived descriptor-based diagnostic; EFL: English as a Foreign Language; ESL: English as a Second Language; PFA: Principal factor analysis; MT: Multiple-trait; MFRM: Multi-faceted Rasch analysis; NS: Native speaker; NNS: Non-native speaker

Acknowledgements

I would like to thank the four raters who participated in this study.

Author's contributions

HJ is the sole contributor of this research paper. The author read and approved the final manuscript.

Funding

The research was funded by Sangmyung University Award Number 2018-A000-0104.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on request.

Competing interests

The author declares that she has no competing interests.

Received: 17 September 2019 Accepted: 8 November 2019
Published online: 30 December 2019

References

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*(3), 371–383.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107. doi: https://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54–74. https://doi.org/10.1080/15434300903464418.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279–293.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29.

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly, 9*(2), 186–203.

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a story retelling speaking test. *Language Assessment Quarterly, 10*(4), 398–422. https://doi.org/10.1080/15434303.2013.824973.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfield, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485–505.

Kim, Y. H. (2010). An argument-based validity inquiry into the Empirically-Derived Descriptor-Based Diagnostic (EDD) assessment in ESL academic writing (unpublished doctoral dissertation). *University of Toronto, Canada*

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304.

Kubota, K. (2018). The potential of empirically derived rating scales for inexperienced raters: A comparative study on rating scales. *JLTA Journal, 21*, 141–159.

Linacre, J. M. (2014). *Facets 3.71 Rasch measurement* [computer software]. Chicago, IL:

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

Jeong, H. (2017a). Narrative and expository genre effects on students, raters, and performance criteria. Assessing Writing, 31, 113–125.

Jeong, H. (2017b, January). EBB Scales in Practice: Challenges and Perspectives. In KATE (Korea Association of Teacher of English)The Proceedings (pp. 316-321).

McNamara, T. (1996). *Measuring second language performance.* London: Longman.

O'Loughlin, K. J. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics, 17*(1), 23–44.

Paas, F., Ayres, P., & Pachman, M. (2008). Assessment of cognitive load in multimedia learning. *Recent Innovations in Educational Technology that Facilitate Student Learning, Information Age Publishing Inc., Charlotte, NC,* , 11-35.

Royal-Dawson, L., & Baird, J. (2009). Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice, 28*(2), 2–8.

Sakyi, A. A. (2003). A study of the holistic scoring behaviors of experienced and novice ESL instructors. Unpublished doctoral dissertation, University of Toronto, Toronto, Canada.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*(2), 163–182.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49–70. https://doi.org/10.2307/3588360.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–3.

Weigle, S. C. (2002). *Assessing writing* Ernst Klett Sprachen.

Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4), 465–492.

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the college English test-spoken English test (CET-SET). *Assessment in Education: Principles, Policy & Practice, 21*(3), 306–325.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.