An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test

Hyunji Park & Xun Yan
University of Illinois at Urbana-Champaign, USA

This two-phased, sequential mixed-methods study investigates how raters are influenced by different rating scales on a college-level English as a second language (ESL) writing placement test. In Phase I, nine certified raters rated 152 essays using a holistic, profile-based scale; in Phase II, they rated 200 essays using a binary, analytic scale developed based on the holistic scale and 100 essays using both rating scales. Ratings were examined both quantitatively through Rasch modeling and qualitatively via think-aloud protocols and semi-structured interviews. Findings from Phase I revealed that, despite satisfactory internal consistency, the raters demonstrated relatively low rater agreement and individual differences in their use of the holistic scale. Findings from Phase II showed that the binary, analytic scale led to much improvement in rater consensus and rater consistency. Another finding from Phase II suggests that the binary, analytic scale helped the raters deconstruct the holistic scale, reducing their cognitive burden. This study represents a creative use of a binary, analytic scale to guide raters through a holistic rating scale. Implications regarding how a rating scale affects rating behavior and performance are discussed.

Key words: many-facet Rasch measurement, holistic scales, analytic scales, writing assessment, rater reliability

Introduction

In performance-based assessment, test scores are influenced by not only the examinee’s ability and task but also the rater and rating scale. In order to make appropriate inferences about test-takers’ true abilities, such effects need to be minimised or at least understood. Much research has focused on rater effects and rating conditions that affect the evaluation of test performance, suggesting that rating is ‘a complex, error-prone cognitive process’ (Cronbach, 1990, p. 584). Raters have been shown to produce different test scores due to
variation in rater severity (Lumley & McNamara, 1995; Wigglesworth, 1993), rating experience (Lim, 2011), rater background (Brown, 1995; Elder, 1993), rating conditions (Ling, Mollaun, & Xi, 2014) as well as how they use different scoring criteria (Cumming, 1990; Cumming, Kantor, & Powers, 2001). While rater training is recommended to reduce such variability among raters (Weigle, 1994, 1998), studies have reported that raters exhibit idiosyncrasy even after training (Vaughan, 1991).

In addition to rater-internal effects, the choice of rating scale also affects how raters rate. In this paper, we define ‘rater behavior’ as raters’ rating processes and how they interact with different aspects of examinee performances and rating scales, and ‘rater performance’ as rater reliability, including both inter- and intra-rater agreement and consistency. We also differentiate three categories of rating scales: holistic, analytic and binary scales. Despite extensive discussions on the strengths and weaknesses of holistic vs. analytic scales in relation to rater reliability and score use, few studies have examined how binary scales influence rater behavior and performance. This study examines how the introduction of a binary, analytic rating scale affects rater behavior and performance.

**Background and Motivation**

While different categorisations of rating scales exist, all rating scales can be largely conceived as either holistic or analytic in nature. Holistic and analytic scales differ in that the former results in a single overall score and the latter in a few subscores and possibly a composite score. A number of scholars favor an analytic scale for the fine-grained score inferences and positive washback while others prefer a holistic scale for practical considerations. In terms of score inferences, analytic scales are arguably more suitable for capturing different learner profiles, which are more reflective of the complex and imbalanced nature in L2 writing development (Hamp-Lyons, 1991, 1995; Jarvis, Grant, Bikowski, & Ferris, 2003; Weigle, 2009; Yan, Kim, & Kotnarowski, forthcoming). They can also offer more diagnostic information and lead to positive washback on teaching and learning (Chan, Inoue, & Taylor, 2015; Hamp-Lyons, 1991, 1995; Knoch, 2011). Despite their advantages, analytic scales are assumed to be time-consuming and expensive to apply (Chan et al., 2015) while it is believed to be faster to train raters to use holistic scales (Klein, Stecher, Shavelson, McCaffrey, Ormseth, Bell, Comfort & Othman, 1998; Weigle, 2009). These practical considerations may explain why holistic scales are more common in writing assessment (Conner, 1991) and recommended for large-scale tests (Jonsson & Svingby, 2007; Xi & Mollaun, 2006).

---

2 Hamp-Lyons (1991, 1995) and some subsequent studies (e.g., Barkaoui, 2007a) use ‘primary trait scoring’ and ‘multiple trait scoring’ in regard to L2 writing assessment instead.
How the two types of rating scales affect rater reliability and rating process is less clear. Holistic scales resulted in higher reliability in some studies (Barkaoui, 2007b, 2011; O’Loughlin, 1994), yet other studies found analytic scales more reliable (Ahmadi, 2013; Song & Caruso, 1996). Bacha (2001), on the other hand, reported that both scales led to high inter-rater reliability and highly correlated scores. Despite such conflicting findings, studies seem to agree that holistic scales may increase the variability in raters’ use of the given rating scale: when rating with a holistic scale, raters seem to use their own rating criteria or place different weights on different criteria (Barkaoui, 2010a; Goulden, 1994; Vaughan, 1991). Accordingly, studies have questioned a high reliability of a holistic scale because a holistic score may disguise such variability among raters (Barkaoui, 2007b, 2011; O’Loughlin, 1994).

Also unclear is the cognitive demand of the two rating scales. Earlier studies argued that holistic scales are cognitively less demanding because raters provide only one overall score (see Jonsson & Svingby, 2007). However, other studies suggested that analytic scales may reduce the cognitive load of rating (Barkaoui, 2007b, 2010b, 2011; Jonsson & Svingby, 2007; Xi & Mollaun, 2006). For example, an analytic scale is found to have ‘training effects’ on less experienced raters in Barkaoui (2011, p. 289) because they could focus on subcategories of the rating scale without having to weigh all aspects of an essay at once (Brookhart, 2018; Cumming, Kantor, & Powers, 2001; Jamieson & Poonpon, 2013).

An alternative rating method that combines the relative strengths of holistic and analytic scales is a binary scale. In this paper, we define binary scales as scales that rely on binary choices or yes/no questions. Binary scales vary in whether the binary choices lead to a holistic score or multiple, analytic scores, so we refer to the former as a ‘binary, holistic scale’ and the latter as a 'binary, analytic scale’.

In the language assessment literature, examples of binary, holistic scales include the empirically derived, binary-choice, boundary-definition (EBB) scale (Upshur & Turner, 1995; Jamieson & Poonpon, 2013) and the Performance Decision Trees (PDTs) (Fulcher, Davidson & Kemp, 2011). By guiding raters through hierarchical binary questions, these scales are assumed to lessen the cognitive burden on raters while providing diagnostic information regarding test takers’ language performance. For example, the PDTs in Fulcher et al. (2011) are designed to provide a holistic score on speaking performances on a test of service encounters. However, they can also provide analytic subscores and diagnostic information on 'Discourse Competence' and 'Pragmatic Competence'. Similarly, Jamieson & Poonpon (2013) adopted the EBB scale to develop analytic ‘rating guides’ for the TOEFL iBT’s integrated speaking tasks to complement the existing holistic scale. The aim of the rating guides was to help students and teachers by providing concrete goals and
diagnostic information and to guide raters to use the holistic scale in a more balanced way.

In contrast, binary, analytic scales may come in the form of a checklist (e.g., Kim, 2011; Struthers, Lapadat, & MacMillan, 2013; Xie, 2017). Checklists consist of a list of binary questions that are grouped under different criteria, and the answers to the binary questions for each criterion are summed to derive a subscore. While EBBs, PDTs and checklists all consist of binary questions (or statements for most checklists), checklists differ from the others in that they provide no clear hierarchy among the questions/statements. For example, the Empirically-derived, Descriptor-based Diagnostic (EDD) checklist in Kim (2011) consisted of 35 descriptors designed to provide five subscores of content fulfilment, organisational effectiveness, grammatical knowledge, vocabulary usage and mechanics.

Binary scoring methods seem to provide a promising alternative to holistic and analytic scoring. First, considering that the presence and absence of a feature is easier to evaluate than its degree, these scales may reduce the cognitive burden of rating (Brooks, 1957). More importantly, these scales may provide guidelines to help raters to conform to the rating criteria, thereby ensuring reliability and reducing raters’ cognitive difficulty (Connor-Linton, 1995; Jamieson & Poonpon, 2013; Williams & Rink, 2003). However, the few studies that have utilised binary scales do not provide a clear answer as to how binary scales influence rater behavior and performance vis-a-vis traditional holistic or analytic scales. Even though Ahmadi (2013) concludes that the binary, analytic checklist resulted in higher rater reliability, Jamieson and Poonpon (2013) and Struthers et al. (2013) report a low inter-rater agreement and a weak internal consistency among raters, respectively. Moreover, the effect of binary scales on raters’ cognitive load is clearly under-studied. As these studies focus on scale development, they only examined rater perception of the scales and the scale development process; however, they did not analyse how rating results and rating process may be affected by the introduction of binary scales.

To address this gap in the literature, the primary goal of this study is to provide empirical evidence as to if a binary scale leads to high rater reliability and helps reduce raters’ cognitive burden in the context of an ESL placement test at a large US university. In so doing, this study adopts a sequential mixed-methods design to provide a fuller picture of the change in rater performance and behavior based on both quantitative and qualitative data. In Phase I, how raters use a holistic scale was analysed. Based on the findings, a binary, analytic scale was developed as a way to deconstruct and complement the existing holistic scale. In Phase II, rater performance on the binary scale and on the holistic scale after the introduction of the binary scale was analysed. Rater performance
measures in the two phases were then compared to address how the introduction of the binary, analytic scale affects rating process and results. This study represents an important and necessary step given the scarcity of literature on how these rating methods affect rater agreement, inter-rater consistency, and rating process.

The current study is similar to Jamieson and Poonpon (2013) in that the binary scale was designed to complement an existing holistic scale. However, it also differs in that the binary questions were used to calculate two subscores and that rater performance with the holistic scale with and without using the binary, analytic scale was analysed. The following research questions were investigated in the study:

1. How do raters perform with a holistic scale in terms of rater reliability and rating process?
2. How do raters perform with a binary scale in regard to rater reliability and rating process?
3. How does rater performance with a holistic scale change after the introduction of a binary scale?

Method

Context for the study

The context for the study is the English Placement Test (EPT), developed and administered at a large US university. International students who do not meet the English requirement set by the university’s mandate are required to take the EPT to receive appropriate ESL writing and pronunciation course placement. The EPT has both a writing and speaking section, but the writing section is the focus of this study. The current version of the writing test is an integrated essay writing task, where students are asked to write an argumentative essay based on information given in a lecture video and a set of reading materials. Student essays are rated by two to three certified EPT raters (for more information about the test, see Cho, 2001; Kim, Bowles, Yan, & Chung, 2018).

EPT essays are rated on an empirically-developed, profile-based rating scale. The holistic scale, developed through a long-term collaboration between ESL teachers and language testers at the university, aims to capture six writer profiles across four proficiency levels, illustrating different combinations of strengths and weaknesses in argument development (ARG) and lexic-grammar (LEX). As illustrated in Figure 1, a well-balanced writer who demonstrates strengths in both of the criteria is given the highest level, A, while a writer with weaknesses in both receives the lowest level, D. Students in
between, who exhibit uneven learner profiles are further divided into four profiles across two proficiency levels (B and C). Students whose argumentation skills are stronger than their lexico-grammar are categorised as B1 or C1 while the opposite cases receive B2 or C2. The profile difference (1 or 2) does not affect placement results but provides additional diagnostic information for stakeholders (see Appendix A for the full scale).

![Figure 1. Visualisation of the profile-based holistic scale](image)

**Participants and materials**

The study involved 452 EPT student essays and nine certified EPT raters. Among the essays, 200 essays were benchmark essays: benchmark essays have been rated unanimously by certified raters and selected by the scale developers to represent each profile of the holistic scale for rater training purposes. The other 252 essays were operational test essays from three test windows. The nine raters were all experienced graduate teaching assistants in the ESL program. Four raters were L1 speakers of English while the rest were highly proficient L2 speakers of English. Most of the raters (n = 7) were pursuing their master’s degree in Teaching English as a Second Language at the time of the study while two were doctoral students studying languages and linguistics. Even though the raters differed in their degree programs and L1/L2 English speaker status, they were assumed to perform similarly as certified EPT raters given their successful completion of the certification program. All raters performed satisfactorily in the recalibration session held just before the first rating session reported in this paper.

**Research Design**

The current study adopted a sequential mixed-methods design, where quantitative data about rater performance were supported by qualitative data. As shown in Figure 2, it consisted of two phases: Phase I investigated the holistic scale’s effects on raters and
Phase II both the binary scale’s effects on raters and its impact on raters’ use of the holistic scale.

**Figure 2. Two-phase, sequential mixed-methods research design**

*Phase I: Rater behavior and performance on the holistic scale*

In Phase I, the nine raters’ ratings on 152 operational test essays were collected for the quantitative analysis of rater performance on the holistic scale. Each essay was randomly assigned to two raters, who received approximately the same number of essays to rate. Third ratings were excluded from data analysis as third raters knew they were rating essays on which two raters disagreed and their ratings may have been affected by that knowledge. The qualitative data of rater performance were collected via semi-structured interviews with four raters (Raters 1, 2, 3, 6), which included verbal protocol/report or think-aloud protocol components. We interviewed the raters individually, and each
interview lasted for about an hour (58-84 minutes). All interviewees signed a consent form and were asked the following questions:

1. Can you describe how you usually rate an essay?

2. Do you experience any difficulties when rating with the holistic scale?

After answering those questions, the raters performed think-aloud protocol as they rated six benchmark essays that they had rated a week before. When needed, the interviewer asked clarification questions and elicited responses. For the last few minutes, the raters were asked to provide feedback on an earlier version of the binary scale. The interviews were recorded and transcribed verbatim for data analysis.

The rating data were analysed using Many-Facet Rasch Measurement (MFRM) with FACETS (Version No. 3.71.4) (Linacre, 2015). For the holistic scale ratings, the following two-facet model was adopted to fit the data (two facets: examinee’s ability and rater severity):

\[
\ln \left( \frac{p_{njk}}{p_{njk-1}} \right) = \theta_n - \alpha_j - \tau_k,
\]

where

- \(p_{njk}\) = probability of examinee n receiving a rating of k from rater j,
- \(p_{njk-1}\) = probability of examinee n receiving a rating of \(k-1\) from rater j,
- \(\theta_n\) = proficiency of examinee n,
- \(\alpha_j\) = severity of rater j,
- \(\tau_k\) = difficulty of receiving a rating of k relative to \(k-1\).

Rater performance was analysed with regard to rater consensus and rater consistency measures produced by the MFRM models. Rater exact agreement percentages estimated by MFRM were used as the rater consensus measure. Intra-rater consistency was investigated using the rater infit statistic. Infit statistics below 0.5 is considered overfit, those between 0.5 and 1.5 acceptable fit and those over 1.5 misfit (Linacre, 2015). A more conservative range suggested for an acceptable fit is between 0.7 and 1.3 (Bond & Fox, 2015). Overfit means rater performance is too predictable and conforming given the Rasch model. Misfit, on the other hand, means raters are inconsistent and unpredictable (Myford & Wolfe, 2003, 2004).
An inductive approach was used to analyse the qualitative data (Cohen, Manion, & Morrison, 2018; Thomas, 2006). The interview transcripts were repeatedly read, coded and analysed to discover emerging themes. After a few iterations and revisions, a coding scheme was developed through discussion and applied to analyse the data (see Appendix B).

Combining the analysis of rater performance on the holistic scale and feedback from the raters in Phase I, we developed a binary, analytic scale (shown in Figure 3), to help raters operationalise the EPT constructs and score the essays more consistently. The binary scale consists of two sets of questions for ARG and LEX. Each component had four questions, and after additional feedback from three raters and two testers, the last questions of both components were given more weight than others. Given the successful implementation of the holistic rating scale and the rater training program over the past three years, the binary, analytic scale was developed based on the holistic scale to provide an alternative format of the same content for raters to ensure rater reliability. The rationale behind the development of the analytic scale was two-fold: 1) a holistic scale developed by a data-driven approach tends to be complex, and 2) an analytic or binary scale developed based on such a holistic scale are thought to provide guidelines to raters to make the rating process cognitively less demanding. As a result, the binary scale reflected the major subcategories of the holistic scale and was designed to focus raters’ attention on each of the major criteria when they rate essays.

### Argumentation (ARG)

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes or No (1 or 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1. Is the writer’s stance clear?</td>
<td></td>
</tr>
<tr>
<td>A2. Do most of the body paragraphs provide relevant evidence?</td>
<td>Yes or No (1 or 0)</td>
</tr>
<tr>
<td>A3. Do most of the body paragraphs provide explanations/analysis of the evidence?</td>
<td>Yes or No (1 or 0)</td>
</tr>
<tr>
<td>A4. Are the explanations/analysis effective?</td>
<td>Yes or No (0 or 2)</td>
</tr>
</tbody>
</table>

### Lexico-Grammar (LEX)

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes or No (1 or 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1. Is it easy to understand most of the essay?</td>
<td></td>
</tr>
<tr>
<td>L2. Are there many minor grammatical errors?</td>
<td>Yes or No (0 or 1)</td>
</tr>
<tr>
<td>L3. Do you notice major grammatical errors?</td>
<td>Yes or No (0 or 1)</td>
</tr>
<tr>
<td>L4. Does the writer often use appropriate vocabulary to express ideas precisely?</td>
<td>Yes or No (2 or 0)</td>
</tr>
</tbody>
</table>

*Figure 3. The binary, analytic scale*
Because the raters expressed difficulties with evaluating students’ language abilities, training materials for the analytic scale were also created with an emphasis on the LEX criteria. The LEX criteria reflected the raters’ comments on lexico-grammatical errors. The training materials consisted of detailed explanations of the binary questions and annotated benchmark essays. The raters had a week to review the materials before the following training session. Then raters were given a group training session. Considering the raters’ extensive training and experience with the holistic scale, the training session for the binary, analytic scale was limited to one hour.

Phase II: Rater performance on the binary scale and the holistic scale.

In Phase II, ratings on 200 benchmark essays and 100 operational test essays were collected for quantitative analysis. Each benchmark essay was rated by five raters (1000 ratings on 200 benchmark essays) while each operational essay was rated by three raters (300 ratings on 100 operational test essays). The number of raters per essay was determined to maximise connections among raters while keeping the regular essay rating schedules. The raters rated the essays in four rounds over a month. They were given about four to seven days to complete one rating round and a break for a few days in between. This rating schedule was similar to regular rating schedules.

As shown in Figure 4, this study analysed a total of five rating sessions—one in Phase I and four in Phase II. The first rating session (H1) was conducted using only the holistic scale to answer the first research question (RQ1). In Phase II, ratings with the binary scale over four rating sessions were collected to investigate the second research question (RQ2). Among them, two rating sessions on benchmark essays (B1, B2) were rated using only the binary scale. The other two rating sessions on operational test essays (HB1, HB2) were rated using both the holistic and binary scales to investigate the relationship between the two scoring methods. In order to answer the third research question (RQ3) on the binary scale’s effect on rater performance with the holistic scale, rater performance of H1 was compared with that of HB1 and HB2. Although the raters did not rate the same essays across sessions, the comparison was made on the assumption that we had observed a similar distribution of student ability and rater performance per rating session over the past few years. Raters 6 and 7 were unable to join the last rating session (HB2) due to schedule conflicts, so seven raters rated HB2 while all the other rating sessions were rated by the nine raters. Table 1 summarises the rating sessions of this paper in terms of the number of raters, essays and raters per essay and the total number of ratings per rating round.
Similar to Phase I, the quantitative data were complemented by the qualitative data. The qualitative data of Phase II consisted of a semi-structured, retrospective interview held after the first rating round and a post-rating survey. As for the interview, the nine raters attended a group meeting for 84 minutes to provide feedback on the binary scale. The raters signed a consent form, and the group interview was recorded and transcribed. The following questions were asked, and follow-up questions and comments were discussed:

1. Did you experience any difficulties when rating with the binary scale?

2. Did you have any questions or feedback on the binary scale?

After all rating rounds were complete, raters completed a questionnaire to provide feedback on both the holistic and binary scales. However, the survey results were quite similar to the interview results and therefore not included in this paper for lack of space.

The data analysis in Phase II was similar to that in Phase I. Rater performance with the holistic scale was analysed using the same two-facet Rasch model. In addition, the binary scale ratings were analysed by fitting a model for the ARG questions and LEX questions separately in order to obtain different scores for the subcategories. This model had an additional facet to estimate, the difficulty of the binary scale questions. The interview
data of Phase II were also analysed using an inductive approach and iterative coding through discussion (see Appendix C for the resulting coding scheme).

Results

RQ1: Rater behavior and performance with the holistic scale

Rater performance statistics on the holistic scale are summarised in Table 2. In terms of internal consistency, the infit statistics showed that all raters performed satisfactorily on the holistic scale, falling in the more conservative range of 0.7 and 1.3. The raters’ standardised fit statistics were also acceptable (standardised fit < 2). The exact agreement measure ranged from 15.2% to 43.8% with the average agreement of 34.37%. Raters 3, 2 and 8 showed the highest agreement statistics (over or around 40%) while Raters 4 and 1 had the lowest (below 30%). The analysis of rater reliability with the holistic scale showed that the raters rated the essays consistently, but their agreement statistics left much room for improvement even after intensive rater training (see Appendix D for further rater-related statistics).

Table 2. Rater reliability measures with the holistic scale

<table>
<thead>
<tr>
<th>Rater ID</th>
<th>Infit</th>
<th>Standardised fit</th>
<th>Exact Agreement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>1.25</td>
<td>0.8</td>
<td>25.8</td>
</tr>
<tr>
<td>Rater 2</td>
<td>0.9</td>
<td>-0.2</td>
<td>40</td>
</tr>
<tr>
<td>Rater 3</td>
<td>0.99</td>
<td>0.0</td>
<td>43.8</td>
</tr>
<tr>
<td>Rater 4</td>
<td>0.94</td>
<td>0.0</td>
<td>15.2</td>
</tr>
<tr>
<td>Rater 5</td>
<td>0.83</td>
<td>-0.7</td>
<td>36.4</td>
</tr>
<tr>
<td>Rater 6</td>
<td>0.81</td>
<td>-0.7</td>
<td>35.3</td>
</tr>
<tr>
<td>Rater 7</td>
<td>1.05</td>
<td>0.2</td>
<td>36.7</td>
</tr>
<tr>
<td>Rater 8</td>
<td>1.19</td>
<td>0.8</td>
<td>40.6</td>
</tr>
<tr>
<td>Rater 9</td>
<td>1.05</td>
<td>0.2</td>
<td>35.5</td>
</tr>
</tbody>
</table>

During the think-aloud protocols, it was found that the raters showed much variability in their use of the holistic scale. It was most striking that some raters considered criteria that were not specified on the holistic scale even after successfully completing an intensive rater certification program. Rater 2 focused on the presence of a counter-argument and Rater 3 on the parallel structure of topic sentences, the quality of the introduction paragraph and the usage of anecdotal evidence even though those were either de-emphasised or absent in the holistic scale. Below are a few think-aloud excerpts that illustrate their appeal to scale-irrelevant criteria.

Rater 2: If I see that any words that signal a counter-argument, just because that shows, you know, a good argument. (code: 3-i)
Rater 3: *If topic sentences are parallel, [it’s good].* (code: 3-c)

Rater 3: *If the introduction says something random, I know they are not going to be great in the body paragraphs.* (code: 3-a)

Rater 3: *If I see that they’re using their anecdotal evidence, I’ll read them in more details.* (code: 3-a)

Also noticeable was that raters often focused only on the ARG criterion because they felt rating the LEX criterion was more difficult. This practice is problematic because raters were not fully utilising the holistic scale and thus potentially undermining the validity and reliability of test scores. In fact, the raters rarely made comments on LEX as they rated the essays and had to be prompted to comment on the LEX components. As the following excerpts illustrate, they did not consider LEX as much because 1) they found the criterion more difficult to judge, 2) they were too used to ESL learners’ mistakes to notice lexico-grammatical issues, and 3) the LEX criterion was de-emphasised in the curriculum.

Rater 1: *After reading so many essays, I understand everything!* (codes: 2-a, 4-a)

Rater 2: *Lexico-grammar is difficult because I understand 80% or higher.* (codes: 2-a, 4-a)

Rater 6: *I don’t notice the language issues as much. In my brain, I just skim over them I don’t see teeny-tiny grammar issues. […] I don’t want to overthink this. I want to go with my gut.* (code: 2-a)

Rater 2: *The ESL writing courses are more about argumentation or rather organisation, so I tend to pay more attention to argumentation than language.* (codes: 2-b, 5-a)

That said, when asked, the raters seemed to agree on what kinds of lexico-grammatical features needed further attention. All raters mentioned that common lexico-grammatical errors found in student essays such as article errors tend not to impede meaning, and that penalising such errors would be counter-productive. Instead, many commented that overly frequent errors and errors that require readers to reorganise sentences were more difficult to fix and needed more attention.

Rater 6: *If we get obsessed with, for example, article issues, not many will get a good grade […] because article issues are so prevalent […] even a very good writer makes a lot of them.* (codes: 4-a, 4-b, 4-c)
Rater 2: [Article errors are] a minor issue because it’s not going to impede my understanding. […] [Errors like] run-on sentences confuse me more. (codes: 4-a, 4-b)

Rater 3: If students write this long sentence, long in a bad way, […] it can be difficult to follow. (codes: 4-a, 4-b)

Rater 1: Sometimes they have verb tense errors […] and pluralise nouns, which do not exist. But those are minor errors. […] some sentences with syntactic errors just do not make any sense. (codes: 4-a, 4-b, 4-c)

Finally, the raters sometimes rated based on their teaching experience or sympathy towards the test takers instead of rating the essays according to the holistic scale. For example, Rater 6 rated an essay higher than others because she thought the student’s logic was good enough as compared to the students in her class, not basing her decision on the rating scale. In another example, Rater 1 assigned a higher score to an essay even though it failed to answer the essay question. In the interview, she said she did not want to rate the essay according to the rating scale as she felt the student could have written a good essay if they had not misunderstood the prompt.

Rater 6: I feel like this author is pretty close to writing a really good essay. (code: 5-b, 5-c)

Rater 1: I know I’m supposed to give a D to this. [But] it’s just that the student didn’t get the directions well in the test settings […] It’s not fair for me to put them in two courses. (code: 3-b, 5-c)

These comments seem to suggest that raters often resort to evaluating the essays based on their own set of criteria as an ESL teacher. Their commitment to the ESL program and teaching may be admirable, but this unfortunately meant that the raters were not using the scale as they were supposed to and needed more guidelines on how to use the rating scale.

The variability in the use of the holistic rating scale among the raters observed in the interviews may explain why the rater agreement statistics were low with the operational test essays. All raters were internally consistent because they considered a similar set of criteria each time they rated essays, but their alignment with each other was not ideal because each rater relied on their own idiosyncratic criteria along with those specified on the holistic scale. The quantitative and qualitative analyses combined suggest that rater performance, while internally consistent, has room for further improvement in terms of inter-rater agreement. The rating scale might benefit from further revision in order to
help raters better operationalise the constructs, especially with respect to the rating of lexico-grammar. These findings formed the empirical basis for the development of the binary, analytic scale (see Figure 3) to complement the holistic scale, along with more training materials to exemplify the rating of lexico-grammar.

**RQ2: Rater performance with the binary, analytic scale**

Figure 5 shows how the rater consistency measure changed over the rating sessions with the binary scale. The box plot on the left shows the range of infit statistics among the raters with the average value shown as a line within each box while the line graph on the right represents the median value of the group. Although there was an outlier rater in HB1 and HB2, the raters showed self-consistency in the acceptable range with the analytic scale (infit statistics 0.7-1.3; standardised fit statistics < 2). Overall, ARG scores showed lower infit statistics than LEX scores, suggesting that the raters were slightly more consistent with the ARG criteria.

![Figure 5. Rater consistency with the binary, analytic scale](image)

Figure 6 summarises raters’ exact agreement (%) when they used the binary, analytic scale. Again, the box plot shows the average and spread while the line plot shows the median values across the rating rounds. As shown in the box plot and line chart, the raters showed high exact agreement in terms of both the ARG and LEX criteria throughout the four rating rounds (80.53% and 70.18% respectively). The LEX criteria showed a consistently lower agreement than the ARG criteria, suggesting that raters might need more training with the LEX criteria and/or that the LEX criteria need revisions. However,
another explanation could be that LEX as specified in the assessment context might be a more difficult criterion to agree on than ARG.

![Figure 6. Rater agreement with the binary, analytic scale](image)

Rater reliability measures with the binary scale show the raters’ appropriate internal consistency and high inter-rater consensus with the binary scale. The raters performed consistently well throughout the four rating rounds even though they received a very short training with the new scale. Their ratings were more reliable on the ARG questions than the LEX ones.

After the first round of ratings with the binary scale, a group meeting was held to provide additional help and collect feedback on the binary scale. Even though the quantitative analysis showed positive results overall, the qualitative data analysis shows that the binary scale was received differently by different raters. Rater 2 said the analytic scale was more difficult because she was used to rating with the holistic scale. Rater 1 and 8 were more neutral about the two scales, commenting on both the advantages and disadvantages of the binary scale.

**Rater 2:** I think [the analytic scale was] little more difficult just because I’m not familiar with it. (code: II-B)

**Rater 5:** The holistic scale is easier because you can be in charge of what to consider and how to weigh them. I kept having to re-read the definitions [...] [because] I’m now used to the current scale. (code: II-B)

**Rater 8:** I didn’t think it was more difficult to use than the other one. [...] But I think we need more questions to consider. (codes: II-C, III-A)
Interestingly, these three raters have shown a higher rater reliability during the rater training program and were considered very reliable raters among the group. Their confidence and familiarity with the holistic scale may explain why they found using the analytic scale more difficult than others. The rest of the raters said they liked the analytic scale better as they found it 'easier' and 'more objective'.

Rater 1: *It was much easier.* (code: II-A)

Rater 6: *It felt more objective to me. I feel a lot less unsure about the ratings I was giving.*  
*I personally really love it.* (code: II-A)

All in all, the raters performed consistently better with the binary scale both in terms of rater consensus and consistency. Both measures were better on ARG than LEX, suggesting the relative difficulty of the LEX criterion, even though the measures were much improved with the adoption of the binary scale. The qualitative data analysis suggests that those who felt confident with the holistic scale resisted the analytic scale and found it difficult to adjust to the new rating scale. However, a majority of raters showed positive attitudes toward the analytic scale and said that the analytic scale provided clearer guidelines to the rating process.

**RQ3: Rater performance with the holistic scale after the introduction of the binary scale**

After the binary scale was introduced, the raters were asked to use both the binary scale and holistic scale to provide two sets of ratings per essay. To use the analytic scale as a rating guide, the raters answered the binary, analytic scale questions first and then provided the profile of each essay based on the holistic scale. Figure 7 shows the rater consistency measure with the holistic scale before and after the introduction of the binary, analytic scale. H1 was rated using the holistic scale only while HB1 and HB2 were rated using both rating scales. Even though there was individual variability, a clear downward trend was found in the median value of infit statistic, suggesting that the raters as a group displayed higher intra-rater consistency on the holistic scale after the binary, analytic rating guide was introduced.

We also compared the individual raters’ infit statistics before and after the introduction of the binary scale as shown in Table 3. Even though Raters 3 through 5 recorded a higher infit or less predictability after the introduction of the binary scale, the rest of the raters saw a reduction in their infit statistics.
The binary scale also had positive effects on rater consensus. Figure 8 summarises how exact agreement changed over time on the holistic scale. The median value of the exact agreement on the right increased considerably between H1 and HB1 even though it dropped slightly from HB1 to HB2. A closer look at individual raters’ agreement statistics as shown in Table 4 also confirmed that the binary scale had a positive effect on most of
the raters’ agreement on the holistic scale. Even though Rater 9’s agreement slightly decreased, the rest of the raters showed improvement in rater consensus.

From these comparisons, it is safe to conclude that the introduction of the analytic scale not only enabled the raters to evaluate the lexico-grammar and argumentation criteria more consistently. This positive effect also carried over to their performance on the holistic scale. The majority of the raters benefited from having the extra guidelines and showed improvement in the rater reliability measures.

![Figure 8. Exact agreement with the holistic scale](image)

<table>
<thead>
<tr>
<th>Rater ID</th>
<th>H1 (without the binary scale)</th>
<th>HB1 and HB2 combined (with the binary scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>25.8</td>
<td>27.6</td>
</tr>
<tr>
<td>Rater 2</td>
<td>40</td>
<td>53.9</td>
</tr>
<tr>
<td>Rater 3</td>
<td>43.8</td>
<td>46.5</td>
</tr>
<tr>
<td>Rater 4</td>
<td>15.2</td>
<td>43.1</td>
</tr>
<tr>
<td>Rater 5</td>
<td>36.4</td>
<td>56.2</td>
</tr>
<tr>
<td>Rater 6</td>
<td>35.3</td>
<td>61.1</td>
</tr>
<tr>
<td>Rater 7</td>
<td>36.7</td>
<td>58.8</td>
</tr>
<tr>
<td>Rater 8</td>
<td>40.6</td>
<td>42.9</td>
</tr>
<tr>
<td>Rater 9</td>
<td>35.5</td>
<td>32.9</td>
</tr>
<tr>
<td>Range</td>
<td>(15.2, 43.8)</td>
<td>(27.6, 58.8)</td>
</tr>
<tr>
<td>Average</td>
<td>34.37</td>
<td>47</td>
</tr>
<tr>
<td>SD</td>
<td>8.75</td>
<td>11.61</td>
</tr>
</tbody>
</table>
Discussion and Implications

This study reports on how the introduction of a binary, analytic scale affects rater performance when compared with their use of a holistic scale. The results show that the nine certified raters performed with internal consistency on the holistic scale but showed relatively low inter-rater reliability. In contrast, the binary, analytic scale developed based on the holistic scale resulted in higher rater consistency and consensus when used in conjunction with the holistic scale as well as when used alone. Findings of this study suggest the potential of a binary, analytic scale to reduce rater cognitive pressure and guide raters through the holistic scale, thereby enhancing rater reliability.

The findings in Phase I point to two major challenges to using a holistic scale. First, raters’ high reliability on holistic scores alone may conceal their variability in rating behavior as suggested in previous studies (Barkaoui, 2007b, 2011; O’Loughlin, 1994). The semi-structured interviews with think-aloud protocol components revealed that the raters used the holistic scale differently from one another. Such differences in raters’ actual use of the rating scale may explain the lower rater consensus on operational test essays in the subsequent rating session (H1) with an average exact agreement at 34.37%. Thus, there is clearly a need for investigating rating process in addition to measuring rater reliability when evaluating rater performance with a holistic scale; a high rater agreement statistic on the overall scores alone does not necessarily provide an adequate picture of the raters’ use of the rating scale. Therefore, when using a holistic scale, rater training and rater performance analysis need to take extra care in aligning raters’ use of the scale as well as their final ratings. That said, we are not equating individual rater differences to unreliability. It is impossible to achieve complete interchangeability in terms of scoring among raters (Eckes, 2011). However, the psychological nature of human scoring should not be used as an excuse to avoid addressing rater variability in the use of criteria that are irrelevant to the rating scale. The rating scale provides a common ground for raters to recognise differences, explore scoring strategies, and eventually discover alignment across one another despite individual differences. It is through this interactive and communicative process that rater reliability is achieved. Second, holistic scales, if not clearly specified, leave the actual use of the rating scale much to raters’ discretion, and raters often choose their own operationalisation of the rating scale to reduce the cognitive demand of rating. When using a holistic scale with analytic components, raters are left with the task of considering and weighing different criteria to derive a single score. When faced with such a challenging cognitive task, raters might rely on criteria that are easier for them to operationalise. They might also resort to prior teaching and assessment experience rather than the scale to reduce the cognitive demand of scoring writing performances. If raters rely on their experiences, disagreement might increase especially
when raters differ in their previous experiences such as their language backgrounds, rating experiences and reading strategies. In this regard, the findings of this study confirm the rater effects related to rater background and type found in previous research (Brown, 1993; Cumming, 1990; Cumming et al., 2001; Elder, 1993; Lim, 2011; Ling et al., 2014; Lumley & McNamara, 1995; Vaughan, 1991; Wigglesworth, 1993). Rater training can provide an effective way to reduce the cognitive demand of rating and keep raters from resorting to external criteria (Lim, 2011; Weigle, 1994, 1998). However, the cognitive burden of rating can also be reduced by simplifying the rating process.

The results of Phase II lend strong support to the development and use of a binary, analytic scale for that purpose in L2 writing assessment. It is speculated that the binary, analytic scale led to such improvement in rater performance because it helped raters decompose a given holistic scale into more manageable, simpler sub-components. The scale allows raters to consider a question at a time without having to figure out how to weigh different criteria. Accordingly, this type of scale can also provide additional information about rater performance on the sub-categories. Despite such benefits of a binary, analytic scale, replacing an existing holistic scale with a binary scale may be challenging in an operational test setting due to constraints such as lack of time and resources and resistance from stakeholders (e.g. experienced raters resisting a new rating scale). If so, complementing a holistic scale with a binary, analytic scale can provide an easier option to improve rater reliability. The inquiry into RQ3 showed that the use of the binary, analytic scale in conjunction with the holistic scale led to improved overall rater performance with the holistic scale. While some raters did not benefit as much from the addition of the binary scale, most raters’ improved performance provided support for using a binary, analytic scale as a rating guide for a holistic scale. This finding is in line with previous studies that argued an analytic scale or binary scale would provide additional guidelines for novice raters (Barkaoui, 2010b; Jamieson & Poonpon, 2013; Xi & Mollaun, 2006). In this sense, using a binary, analytic scale to complement the holistic scale provides an opportunity to combine the best of both worlds. By providing the binary, analytic scale along with the holistic scale, the current rating method provides additional guidelines and focuses raters’ attention on each of the major criteria of the holistic scale. With the addition of the binary, analytic scale, the scale will be able to provide diagnostic information about learner profiles, which are essential in L2 writing.

---

3 We acknowledge that the improvement in rater performance after the introduction of the binary scale may be due to more rating experience over time, as novice raters’ performance tends to improve as they become more familiar with the scale and test (c.f. Lim, 2011). However, in our study, the raters had already been familiar with the test and rating. In addition, our regular rater performance analysis over the past three years has shown little change in rater performance over time, so we speculate that the improvement is largely due to the introduction of the binary scale.

Limitations and Future Research

The current study is not without limitations. In particular, we acknowledge that raters’ actual use of rating scales deserved more exploration. While the triangulation of think-aloud protocols, reflective interviews and Rasch measures provided a glimpse into rater process and rater cognition, more think-aloud protocols of raters using the binary, analytic scale alone and as a rating guide to the holistic rating scale would have provided more insights into how raters interact with different rating scales. This also calls for future research into other novel ways to assess rater process and rater cognition to better understand how raters use a rating scale to reach final scores. It would be also interesting to investigate how different characteristics of raters (e.g. L1, rating experience) interact with different types of rating scales in terms of rater cognition and performance. Even though we have mentioned a few instances of individual differences while reporting the quantitative data analysis, how different raters interact with different rating scales deserve more exploration and explanation. Another line of inquiry deserving further exploration is the relative difficulty of LEX criteria as opposed to ARG. Throughout the study, the raters showed less reliability in evaluating students’ language than their rhetorical skills. This could be an issue resulting from the local context of the ESL program, but it is also possible that rater judgement of linguistic skills may be inherently more difficult than that of structural aspects in the context of L2 writing assessment. Future research may shed light on how raters evaluate the content and structure of student essays and how those involve different rating processes. It should also be noted that the current paper did not include the rating scale as one of the facets in the Rasch analysis. Further inquiries into this issue while including the rating scale as a facet may result in interesting insights. Finally, this study provides an example of using Rasch modeling to conduct fine-grained assessment of rater performance. While Rasch modeling is effectively used for quality control purposes, it is also possible to use the method to look into how rater performance changes over time and corroborate it with other types of evidence (e.g. think-aloud protocols) to triangulate rater behavior and more.

Conclusion

This paper investigated how rater behavior is influenced by a holistic scale versus a binary, analytic scale on a college-level ESL writing placement test. Both quantitative and
qualitative analyses of rater performance suggest that the binary, analytic scale reduces the cognitive burden and enhances the reliability of raters by decomposing the evaluation criteria into more concrete, binary choices. The findings of this study provide support for the promise and flexibility of binary, analytic scales to be used either alone or as a rating guide to complement holistic scales in writing assessment. These scales can also provide a cognitive lens into rater behavior in the scoring of writing performances.

References


### Appendix A. Profile-based holistic scale

<table>
<thead>
<tr>
<th>Argument Development</th>
<th>Writing Performance Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong> Strong in both</td>
<td>Essay provides a well-developed argument with a clear controlling idea and supporting details. The argumentation often shows originality.</td>
</tr>
<tr>
<td><strong>B1</strong> Strong in argumentation</td>
<td>Essay provides a sufficiently-developed argument with a clear controlling idea and supporting details.</td>
</tr>
<tr>
<td><strong>B2</strong> Strong in lexicogrammar</td>
<td>Essay provides a sufficiently-developed argument overall with a clear controlling idea and supporting details.</td>
</tr>
<tr>
<td><strong>C1</strong> Weak in lexicogrammar</td>
<td>Essay provides a clear argument overall with supporting details.</td>
</tr>
<tr>
<td><strong>C2</strong> Weak in argumentation</td>
<td>Essay may not provide a clear argument in support of a controlling idea.</td>
</tr>
<tr>
<td><strong>D</strong> Weak in both</td>
<td>Essay lacks complex lexicogrammar and is sometimes vague and/or unclear, but the ideas are easy to understand overall. - Essay has frequent errors that often lead to comprehension/processing difficulty.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paragraph-level Argumentation</th>
<th>Writing Performance Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Paragraph-level argumentation has a well-developed TEA or PIE structure with ample and relevant supporting details. - The supporting details are interconnected to each other.</td>
<td>- Paragraph-level argumentation has a TEA or PIE structure with relevant supporting details but often needs further development. - The supporting details are not consistently interconnected to each other.</td>
</tr>
<tr>
<td>- Essay displays a good command of a wide range of lexicogrammar that conveys meaning clearly and precisely. - Essay has noticeable errors that sometimes cause lapse of clarity or precision.</td>
<td>- Essay displays complex lexicogrammar that conveys overall meaning clearly and precisely. - Essay has minor errors that are not noticeable.</td>
</tr>
<tr>
<td>- Essay lacks complex lexicogrammar and is sometimes vague and/or unclear, but the ideas are easy to understand overall. - Essay has noticeable errors that sometimes cause processing difficulty.</td>
<td>- Essay lacks complex lexicogrammar and is sometimes vague and/or unclear, but the ideas are easy to understand overall. - Essay has noticeable errors that do not impede processing.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lexico-Grammatical Features</th>
<th>Writing Performance Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Essay displays sophisticated lexicogrammar that is characteristic of academic written discourse and conveys finer shades of meaning precisely. - Essay has no or very minor lexicogrammatical errors.</td>
<td>- Paragraph-level argumentation severely lacks supporting details. Even if they are present, they poorly support the point. - The supporting details are not interconnected to each other and often do not advance the controlling idea.</td>
</tr>
<tr>
<td>- Essay displays a good command of a wide range of lexicogrammar that conveys meaning clearly and precisely. - Essay has noticeable errors that sometimes cause lapse of clarity or precision.</td>
<td>- Essay lacks complex lexicogrammar and is sometimes vague and/or unclear, but the ideas are easy to understand overall. - Essay has noticeable errors that do not impede processing.</td>
</tr>
<tr>
<td>- Essay lacks complex lexicogrammar and is sometimes vague and/or unclear, but the ideas are easy to understand overall. - Essay has frequent errors that often lead to comprehension/processing difficulty.</td>
<td>- Essay is often vague and/or unclear to cause much comprehension/processing difficulty. - Essay has frequent errors, which often leads to comprehension/processing difficulty.</td>
</tr>
</tbody>
</table>
### Appendix B.
Coding scheme for the semi-structured interviews in Phase I

<table>
<thead>
<tr>
<th>Categories</th>
<th>Sub-categories</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Overall rating procedure</td>
<td>a. Reading strategies</td>
<td>1-a</td>
</tr>
<tr>
<td></td>
<td>b. Rating strategies</td>
<td>1-b</td>
</tr>
<tr>
<td>2. Difficulties in rating</td>
<td>a. Difficulty in noticing lexico-grammatical issues</td>
<td>2-a</td>
</tr>
<tr>
<td></td>
<td>b. Difficulty in distinguishing profiles (weighing)</td>
<td>2-b</td>
</tr>
<tr>
<td></td>
<td>c. Difficulty in distinguishing levels</td>
<td>2-c</td>
</tr>
<tr>
<td>3. Argumentation criteria</td>
<td>a. Introduction</td>
<td>3-a</td>
</tr>
<tr>
<td></td>
<td>b. Thesis statement/ Central claim</td>
<td>3-b</td>
</tr>
<tr>
<td></td>
<td>c. Topic sentences</td>
<td>3-c</td>
</tr>
<tr>
<td></td>
<td>d. Interconnectedness/ connections/ coherence</td>
<td>3-d</td>
</tr>
<tr>
<td></td>
<td>e. PIE/TEA structure, paragraph-level argumentation</td>
<td>3-e</td>
</tr>
<tr>
<td></td>
<td>f. Source use</td>
<td>3-f</td>
</tr>
<tr>
<td></td>
<td>g. Anecdotal evidence; quality of evidence</td>
<td>3-g</td>
</tr>
<tr>
<td></td>
<td>h. Analysis or explanation of sources</td>
<td>3-h</td>
</tr>
<tr>
<td></td>
<td>i. Counter-argument</td>
<td>3-i</td>
</tr>
<tr>
<td></td>
<td>j. Logic, convincing ideas</td>
<td>3-j</td>
</tr>
<tr>
<td></td>
<td>k. References or citation</td>
<td>3-k</td>
</tr>
</tbody>
</table>
4. Lexico-grammar criteria
   a. Comprehensibility/processing difficulty
   b. Minor errors or major errors
   c. Frequency of errors
   d. Academic register
   e. Long sentences that are confusing
   f. Word choice

5. ESL teaching experience
   a. Difference between ESL curriculum and the holistic scale
   b. Classroom experiences
   c. Test consequences

6. Feedback on binary scale
   a. Positive feedback
   b. Clarification
   c. Suggestions

7. Other comments
   a. Reflection on their rating procedure
   b. Comments/questions on the holistic scale
   c. Suggestions for the holistic scale
   d. Suggestions for rater training
   e. Suggestions for recalibration sessions
   f. Suggestions for rating procedure
## Appendix C.
### Coding scheme for the semi-structured group meeting in Phase II

<table>
<thead>
<tr>
<th>Categories</th>
<th>Sub-categories</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Rating time</td>
<td>A. Faster with the decision tree</td>
<td>I-A</td>
</tr>
<tr>
<td></td>
<td>B. The same as the holistic scale</td>
<td>I-B</td>
</tr>
<tr>
<td>II. Rating difficulty</td>
<td>A. Positive toward the decision trees</td>
<td>II-A</td>
</tr>
<tr>
<td></td>
<td>B. Negative toward the decision trees</td>
<td>II-B</td>
</tr>
<tr>
<td></td>
<td>C. Mixed</td>
<td>II-C</td>
</tr>
<tr>
<td>III. Other feedback on the rating</td>
<td>A. More questions to consider</td>
<td>III-A</td>
</tr>
<tr>
<td>guide</td>
<td>B. Clarification questions</td>
<td>III-B</td>
</tr>
<tr>
<td>IV. Others</td>
<td>A. Rating schedules</td>
<td>IV-A</td>
</tr>
<tr>
<td></td>
<td>B. Comments on other tasks</td>
<td>IV-A</td>
</tr>
</tbody>
</table>
## Appendix D. Other rater-related statistics

<table>
<thead>
<tr>
<th></th>
<th>H1</th>
<th>B1</th>
<th>HB1</th>
<th>B2</th>
<th>HB2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater separation</td>
<td>4.33</td>
<td>5.87</td>
<td>5.96</td>
<td>1.33</td>
<td>1.99</td>
</tr>
<tr>
<td>Rater strata</td>
<td>6.10</td>
<td>8.15</td>
<td>8.28</td>
<td>2.10</td>
<td>2.99</td>
</tr>
<tr>
<td>Reliability (not inter-rater)</td>
<td>.95</td>
<td>.97</td>
<td>.97</td>
<td>.64</td>
<td>.80</td>
</tr>
</tbody>
</table>