Rater training for scoring rubrics: Rater-centered bottom-up approach

This article describes a procedure for training second-language writing raters to use scoring rubrics, and presents ideas for practical adaptation or research projects associated with the training procedure.

Imagine a novice language teacher doing the following:

1) understanding exactly what is meant by rubric descriptors such as *some knowledge of subject, adequate range, limited development of thesis, mostly relevant to topic but lacks detail*;
2) looking for features in a second language writer's composition that appear to match these descriptors, and then, upon finding a match; and
(3) rating the content of the composition as *Good*.

These descriptors are examples from a commonly used rubric for assessing second language writing (Jacobs et al., 1981), in which the example descriptors above represent only 5% of all the descriptors that a rater is supposed to be familiar with while reading and rating compositions.

In programs in which graduate teaching assistants (TAs) need to be trained over a short period of time to rate second language writers' essays and make course placement decisions, rater training must ensure that the novice raters quickly become familiar with the descriptors used in a scoring rubric. Regardless of the specific rubrics used, raters are typically expected to demonstrate knowledge of pre-determined descriptors (such as *some knowledge of subject, adequate range, limited development of thesis, mostly relevant to topic but lacks detail*) and perform the task of applying this knowledge consistently. Such expectations may not be met easily. Joe, Harmes, and Hickerson (2011) show, for example, that lack of transparency in rating scale descriptors can be a factor influencing

raters' performance. At the same time, rater-related factors can also add to the challenge in achieving reliable outcomes in rating (see Barkaoui, 2011 for a comprehensive review).

In this article, first, I will briefly review studies on raters and rater training. Then, I will describe a gap in published studies on rater training. Finally, I will introduce a rater training procedure currently implemented at a four-year university in the U.S. Midwest, for two purposes: (1) to encourage language program administrators or language teacher educators to adopt and field-test the procedure in their own programs, and (2) to open up the possibility of empirically testing the procedure using systematically designed research methods.

Are experienced raters consistently better than novice raters?

Among several rater-related factors that could influence raters' interpretation and application of rating scales, raters' experience (novice vs. experienced) seems to be the most frequently researched factor (Barkaoui, 2011; Greer, 2013; Hamp-Lyons, 1989; Harsch & Martin, 2012; Joe et al., 2011; Weigle, 1994). Research findings on this issue are mixed, contrary to what might be assumed (i.e. the more experienced, the more proficient). In fact, the mixed findings might be a function of the complexity in the way raters' experience interacts with non-rater factors. In Barkaoui (2011), for example, with regard to severity in rating, novice raters and experienced raters behaved more similarly when using analytic scales than they did when using holistic scales. In other words, raters' experience impacts ratings differently depending on the type of rating scales used.

The intriguing nature of raters' experience as a factor was documented with great detail in Joe et al. (2011), which explored rater cognition based on data collected through verbal protocols. In this study, eight faculty experts (experienced raters) and eight undergraduate students (inexperienced raters) participated in rating oral speech performances. Both groups were trained to use an analytic scoring rubric, which included 39 features comprising ten competency dimensions relevant to the construct of the speech performances to be evaluated. The study found that inexperienced raters started out paying attention to rubric features more consistently than did experienced raters, who were found to pay attention to construct-irrelevant features (i.e. features not listed in the rubric) at a higher rate than inexperienced raters did. Over time, however, inexperienced raters attended to rubric features less and less while attending to construct-irrelevant features (such as *the use of note cards* or *memorable thesis statement*, which were not included in the rubric) more and more.

Such findings are quite alarming in that not all changes exhibited by raters as they become more experienced seem to be in the expected direction – i.e. experience is generally expected to be a positive factor. Some researchers even suggest that raters should not be selected based on teaching experience as it is not a significant factor (Royal-Dawson & Baird, 2009). Although research findings about rater experience might be mixed, many researchers have emphasized the importance of rater training in enhancing the quality of raters' performance (Greer, 2013; Lovorn & Rezaei, 2011; Weigle, 1994). The following section will review studies on rater training.

What do we know from studies on rater training?

One of the aims in most rater training involves monitoring rater behavior associated with rater-related factors such as experience, rating style, or rating preferences, and then providing feedback accordingly to achieve the ultimate goal of increasing inter-rater reliability (i.e. different raters performing similarly to one another). On the one hand, studies such as Pufpaff, Clarke, and Jones (2015) and Weigle (1994) reported that rater training did not improve inter-rater reliability. On the other hand, Weigle emphasized the point that "rater training cannot make raters into duplicates of each other, but it can make raters more self-consistent" (p. 32). This statement, then, naturally leads to the question as to what factors might contribute to developing rater self-consistency. Although there has not been much research that directly explored this question, several studies have reported the positive effects of rater training on rater performance in different aspects of the rating task.

In Greer (2013), novice raters practiced assessing ESL compositions following a training workbook, which included experienced raters' feedback on the same compositions that the novice raters were evaluating. After the training, the novice raters reported increased confidence in their rating performance. In another study based on a two-month rater training program (Harsch & Martin, 2012), 13 novice raters completed rigorous weekly assignments consisting of tasks commonly included in rater training such as individual practice and group discussion, using over 1700 writing samples (whittled down from an initial set of over 6000 samples). Although the scope of the study is truly impressive, it is the depth of its rater training that makes it rather unique and remarkable. As a part of their weekly assignments, for example, the novice raters

were actively engaged in revising the wordings on the rating scale. In fact, researchers recommend engaging raters in the development of rating scales (Barkaoui, 2010; Stevens & Levi, 2005). Harsch and Martin (2012) concluded that rater agreement increased when a revised rating scale (i.e. revised based on the novice raters' discussion and input during the training period) was used.

Because the task of rating is replete with a myriad of interacting factors that could influence the process and outcome of rating, research-guided rater training may be essential in most contexts. For example, when training raters, feedback should be provided immediately after rating has occurred (Knoch, 2011). Rater training should promote detailed and analytical understanding of the scoring rubric (Lovorn & Rezaei, 2011; Rezaei & Lovorn, 2010). An eye-movement study on raters' use of a scoring rubric showed that even the physical layout of the rubric can affect raters' attention to each category on the rubric (Winke & Lim, 2015). These are just a few examples of published studies that could guide the design of a rater training.

So what is the gap in published studies on rater training?

In most published studies that either directly or indirectly report the outcome of rater training (Barkaoui, 2011; Joe et al., 2011; Knock, 2011; Lovorn & Rezaei, 2011; Pufpaff et al., 2015; Weigle, 1994), it appears common to follow variations of the same approach, categorically speaking, in the way raters (regardless whether novice or experienced) are initially introduced to a rating scale (either holistic or analytic). First of all, surprisingly, many studies (Barkaoui, 2011; Knock, 2011; Lovorn & Rezaei, 2011) do not provide sufficient detail regarding exactly how raters are introduced to the rating

scales selected in their respective studies. Of the studies (Joe et al., 2011; Pufpaff et al, 2015; Weigle, 1994) that do provide some limited information regarding this part of rater training, the common approach seems to be *Present and Clarify/Explain* with respect to the descriptors on the rating scale.

This clearly is an example of what is defined as the hierarchical approach: "passing onto raters a predetermined view on how they are to interpret the scale wordings, using pre-assessed scripts (so called 'master codes') which are not to be discussed but to be accepted and internalised" (Martin & Harsch, 2012, p. 233). On the one hand, researchers have repeatedly identified the difficulty that raters experience when trying to understand rating scale descriptors (Barkaoui, 2010; Greer, 2013; Hamp-Lyons, 1989; Harsch & Martin, 2012; Joe et al., 2011). On the other hand, as described above, variations of a very top-down approach seem to prevail when it comes to training raters to become familiar with the wordings on the rating scale.

From this perspective, Harsch and Martin's (2012) study, as reviewed in the preceding section, may be considered as an exception in that the raters in their study were engaged in a series of in-depth tasks attending to, analyzing, and revising the descriptors on the scale. These researchers also emphasize the importance of "reaching consensus about how to interpret scripts with reference to scale descriptors" (p. 233). As mentioned earlier, the rater training in their study spanned over a two-month period. Most readers would agree that a rater training program like that, while both impressive and exemplary, is anything but feasible in most real contexts. The reality of most rater training is likely

to resemble the two-hour norming session referred to as a typical rater calibration procedure (Weigle, 1994, pp. 7-8).

Here is a question, then, that a concerned program administrator or teacher educator might ask: *Do people use the top-down approach because they are constrained to the typical two-hour calibration procedure (or however many hours it might take but not the luxury of two months)?* Acknowledging the "time- and resource-intensive" nature of their approach, Harsch and Martin (2012, p. 244) recommend realistic adaptations using existing rating scale descriptors (i.e. not necessarily attempting to revise the descriptors as their raters did). So it appears that, although there is a huge gap between a deeply engaging, albeit extremely unfeasible, approach to rater training and a more commonly practiced top-down approach, careful retooling of the top-down approach can help fill this gap. In the next section, I will introduce an authentic example of a rater training procedure which is characterized as a rater-centered bottom-up approach. (Readers can rest assured that this procedure will not require two months to try!)

What does a rater-centered bottom-up rater training procedure look like?

Bottom-up Approach

The rater training procedure described here, as an example of a rater-centered bottom-up approach, has been implemented in an authentic test context. The following provides some background information about the context:

Location	Four-year university in the U.S. Midwest
Purpose	Placement decisions for writing courses in English for Academic Purposes
Raters	Graduate teaching assistants in an MA-TESL program (1^{st}~4^{th} semester)
Target Texts	Academic essays written by second language writers
Scoring Rubric	Locally revised version of Composition Profile by Jacobs et al. (1981)

Training Protocol (Rater-centered Bottom-up Approach)

[I] Individually **[SG]** Small Groups (3~4) **[WG]** Whole Group **[T]** Trainer
M: Materials used at each step (See Note.)

Description of Steps in the Training Protocol	Rationale for Each Step
1. Activating existing knowledge & expectations about academic writing	
[SG] Brainstorm & consolidate existing knowledge & expectations about academic writing **[WG]** Discuss and summarize **M:** Brainstorm sheets	**Step 1** allows each rater to activate existing knowledge; allows each rater to generate his/her own language to describe features of writing; compile & share entire group's ideas.
2. Evaluating a writing sample based on existing knowledge - without any rubric	
[I] Read Essay #1; Write any/all notable features, good & bad, one feature per sticky-note; Place sticky notes in worksheet; Give a holistic score **[SG]** Compare notes placed in individual worksheets; Compare holistic scores **M:** Essay #1; Sticky-notes; Worksheet	**Step 2** allows each rater to apply existing knowledge; allows each rater to notice features in the writing with no constraints; exposes raters to writing features noticed by others.
3. Familiarization with rating scale descriptors	
[I] Read rating scale descriptors and criteria **[SG]** Discuss & help each other understand concepts & terminologies **[WG]** Review & clarify concepts & terminologies **M:** Rating scale descriptors & criteria handout; handout on Content-to-Form continuum in writing	**Step 3** introduces descriptor language to raters; helps raters to conceptually align their own language with descriptor language; helps identify & clarify gaps between rater-generated language and descriptor language.
4. Matching current knowledge with rating scale descriptors	
[SG] Discuss each note on sticky-notes; Transfer & match each sticky-note with descriptors in the Descriptor Handout **[WG]** Discuss & further clarify descriptors based on questions from SGs **M:** Descriptor Handout (one copy for each SG)	**Step 4** allows raters to map their own unconstrained observations onto descriptors; helps identify, discuss, & resolve writing features that are difficult to map onto descriptors; more importantly, helps raters understand descriptors with self-generated concrete examples.
5. Practice using rating scale descriptors without scores	
[I] Read Essay #2; Use the Descriptor Handout to mark relevant descriptors **[SG]** Compare individuals' markings on descriptor handout **M:** Essay #2; Descriptor Handout	**Step 5** allows raters to practice using the descriptors directly without scaffolding (i.e. no self-generated descriptive notes as with Essay #1); allows another chance to focus on the descriptors with no burden to score the essay numerically.
6. Familiarization with the complete version of rating scale (with score indicators)	
[T] Introduce the complete version of the rating scale (with score indicators) and explain **[I]** Based on markings on Descriptor Handout (from Step 5) and using the rating scale (with score indicators), numerically score Essay #2 **[SG]** Compare scores for Essay #2 **M:** Complete version of rating scale	**Step 6** finally exposes raters to the actual rating scale with score indicators; helps raters perform the task of numerical scoring (not exactly the same as mapping observed writing features onto descriptors); helps deal with two different subtasks (i.e. identifying matching descriptors vs. numerical scoring) with more clarity.
7. Practice using the complete version of rating scale with a familiar essay	
[I] RE-read & score Essay #1 using rating scale **[WG]** Discuss the results & rationale of rating scale **M:** Complete version of rating scale with score indicators	**Step 7** allows raters to apply the rating scale in evaluating a familiar writing sample; provides raters with an opportunity to review & re-assess their own initial evaluation of Essay #1 (performed prior to the introduction of the rating scale).

8. Remaining steps in the protocol
There are a few more steps in the protocol, which are beyond the focus of this article. Some of the remaining steps are similar to commonly practiced norming procedure, and some steps are specific to the local test context. [1]

Table 1. Training Protocol

As Table 1 shows, the training protocol follows a rater-centered bottom-up procedure, which affords the raters step-by-step scaffolding to develop an understanding of and the ability to apply the descriptors on the rating scale. The procedure promotes activating existing knowledge and acquiring new knowledge of technical concepts/terminologies through a sequence of small tasks rather than through top-down imposition of abstract descriptors onto the raters. For many novice raters, learning to use a rating scale with pre-determined descriptors includes an element of language acquisition. It is not a mere coincidence that, in many ways, the procedure introduced here resembles language learning activities based on the task-based language teaching (TBLT) approach, in which language acquisition occurs as a natural part of successful completion of communicative tasks (Van den Branden, 2006).

The rater training procedure introduced here also provides scaffolding for one of the subtasks of rating that present a unique challenge for most raters, namely translating descriptors into numerical scores. Studies have shown that both novice raters (Greer, 2013) and experienced raters (Hamp-Lyons, 1989) find this subtask very difficult. In the procedure described in Table 1, raters are assisted to deal with this challenge in two ways: (1) initial steps in the procedure focus on the descriptors without the 'burden' of matching them with numerical scores; and (2) the complete version of the rating scale, a

[1] Due to space limitation, this section has been abridged. Please feel free to contact the author for more information or samples.

locally revised version of the composition profile by Jacobs et al. (1981), presents numerical scores in subsets to match raters' judgments based on descriptors.

Unlike the two-month rater training described in Harsch and Martin (2012), which is quite impressive and ambitious, the training procedure introduced in this article is bundled with realistic and practical advantages:

1. the procedure can be easily adopted and adapted in most rater training contexts;
2. the sequence will work well with any type of rating scales;
3. it only takes 2~4 hours to complete the entire protocol; and finally and more importantly,
4. it does not cost much (i.e. if you can afford lots of sticky-notes).

The last two items directly address "the time- and resource-intensive" challenge of Harsch and Martin's (2012, p. 244) otherwise exemplary rater training model. Hopefully, these practical advantages would encourage many readers of this article to consider employing this training procedure.

Aside from the obvious practical advantages, the most critical advantage of this procedure, at least based on informal observations during several semesters of implementation, involves the change in dynamics and roles between the trainer and the raters in training. As the protocol shows, at each step, raters are actively engaged in small doable tasks either independently or in collaboration with peer raters. Because the steps are sequenced to promote learning-by-doing, the procedure does not require much top-down talk from the trainer.

When this new procedure was first implemented a few semesters ago, raters who had experienced the previous format resembling the *Present and Clarify/Explain* approach enthusiastically commented that the new procedure felt stress-free, engaging,

and helpful. As the trainer in this incidence, I too noticed unexpected changes when first implementing the new procedure. It felt as though I did not have to do anything during the procedure because the raters were doing all the work for themselves!

What more can be done with this training procedure?

Although the rater-centered bottom-up training procedure is strongly recommended, the caveat is that it has not been empirically tested. First, interested readers are encouraged to consider field-testing this procedure in their various test contexts. It can be modified to fit the needs and capabilities of each context. One example might be replacing the use of sticky-notes with a digital/online tool to help raters generate, compile, and compare the features they observe in the writing sample they evaluate. This is actually an attractive idea, which can lead to the next point of this discussion, namely research possibilities.

The hands-on aspect of using sticky-notes is actually a very positive and valuable element of the procedure, and it helps raters ease into the sequence of tasks in the procedure. Its non-digital nature, however, has been an obstacle in converting the notes into analyzable data. These rater-generated notes can reveal interesting aspects of rater cognition. Designing empirical studies to capture such data to learn more about rater cognition would not only benefit the field of language education but also the field of education in general. In fact, in the special issue of the journal *Educational Measurement: Issues and Practice* devoted to rater cognition, Myford (2012) emphasizes that more research on rater cognition is needed.

Other research questions worth exploring including obvious ones such as *Is the bottom-up rater training procedure more effective than the more commonly practiced top-down approach? Is there any difference between the two approaches in improving rater self-consistency (emphasized as the main benefit of rater training)?* These are just a few examples, and readers are encouraged to pursue their own research questions associated with the rater-centered bottom-up rater training procedure introduced in this article.

This article started by asking readers to imagine a novice rater performing a presumably learnable, but indeed tremendously challenging, task of using a rating scale to make decisions about learners' proficiency. Anecdotal evidence and informal observations suggest that rater training, for both novice and experienced raters, need not be like that – top-down, opaque, and anxiety-inducing. Instead, a rater-centered bottom-up approach can make the process more transparent and positively engaging. However, for this statement to be generalizable, we need empirical evidence, and this research topic is open to any interested readers.

References

Barkaoui, K. (2011). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*, 31057.

Greer, B. (2013). *Assisting novice raters in addressing the in-between scores when rating writing*. (Master's thesis). Retrieved from BYU ScholarsArchive.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. W. Dechert & Raupauch (Eds.), *Interlingual processes* (pp. 229-244). Tubingen: Buner Narr.

Harsch, C. & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*, 228-250.

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

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: a mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice, 18*, 239-258.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing, 28*, 179-200.

Lovorn, M. G. & Rezaei, A. R. (2011). Assessing the assessment: Rubrics training for pre-service and new in-service teachers. *Practical Assessment, Research & Evaluation, 16*, 1-18.

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice, 31*, 48-49.

Pufpaff, L. A., Clarke, L., & Jones, R. E. (2015). The effects of rater training on inter-rater agreement. *Mid-Western Educational Researcher, 27*, 117-141.

Rezaei, A. R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*, 18-39.

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

Stevens, D. D. & Levi, A. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning.* Virginia: Stylus Publishing, LLC.

Van den Branden, K. (Ed.). (2006). *Task-based language education: From theory to practice.* Cambridge, UK: Cambridge University Press.

Weigle, S. C. (1994). Using FACETS to model rater training effects. Paper presented at the *Language Testing Research Colloquium* (Washington, DC).

Winke, P. & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing, 25*, 38-54.

Author Information

Choonkyong Kim
ckim@stcloudstate.edu
Choonkyong Kim is a faculty member at St. Cloud State University. She teaches courses in applied linguistics in the TESL Program and directs the English for Academic Purposes Program.