Language Assessment Quarterly, 9: 270–292, 2012

ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434303.2011.649381

Routledge
Taylor & Francis Group

Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior

Thomas Eckes
University of Bochum

This research investigated the relation between rater cognition and rater behavior, focusing on differential severity/leniency regarding criteria. Participants comprised a sample of 18 raters examined in a previous rater cognition study (Eckes, 2008b). These raters, who were known to differ widely in their perceptions of criterion importance, provided ratings of live examinee writing performance. Based on these ratings, criterion-related bias measures were estimated using many-facet Rasch measurement. A cluster analysis of bias measures yielded four operational rater types. Each type was characterized by a distinct pattern of differentially severe or lenient ratings on particular criteria. The observed bias patterns were related to differential perceptions of criterion importance: Criteria perceived as highly important were more closely associated with severe ratings, and criteria perceived as less important were more closely associated with lenient ratings. Implications of the demonstrated link between rater cognition and rater behavior for future research into the nature of rater bias are discussed.

Recent increase in the use of constructed-response tasks in high-stakes language testing and assessment has led to an increasing need to understand how humans evaluate examinees' responses (G. T. L. Brown, 2010; Engelhard, 2002; McNamara, 1996; Wigglesworth, 2008). Being human, these evaluations inevitably are subject to a wide range of factors that threaten the validity and fairness of the assessment outcomes. Figuring prominently among these are factors referring to rater characteristics such as rater severity or leniency, to the perception of scoring criteria, and to the complex nature of the rating process itself (Bejar, Williamson, & Mislevy, 2006; Lumley, 2005; Wolfe, 1997). In particular, evaluating an examinee's written or spoken performance is a challenging task that in most situations puts high cognitive demands on raters (Hamp-Lyons & Henning, 1991; Luoma, 2004). The focus of the present research is on how raters of written performance perceive scoring criteria and how this perception is associated with criterion use in an operational assessment context.

As reviewed in the next two sections, there has been a considerable number of studies addressing various aspects of rater cognition involved in performance evaluations (e.g., differential weighting of criteria or selective attention to performance features), as well as research into the kind of decision-making behaviors and biased use of scoring criteria in rating examinee performance, but research directed at investigating the link between rater cognition and rater behavior

Correspondence should be sent to Thomas Eckes, TestDaF Institute, University of Bochum, Massenbergstr. 13 b, 44787 Bochum, Germany. E-mail: thomas.eckes@testdaf.de

has been scarce. Probing into this link may serve to inform rater training and rater monitoring activities and, thus, enhance the overall quality of rater-mediated assessments. Adopting a classificatory approach to rater variability, the following study extends earlier work on rater types characterized by distinct perceptions of scoring criteria (Eckes, 2008b, 2009b) to the question of how raters belonging to different types actually make use of criteria in an operational scoring session. In particular, this study relates differential perceptions of criterion importance to patterns of unduly harsh or lenient ratings awarded to examinees. The basic assumption is that perception and interpretation of scoring criteria play a key role in accounting for criterion-related rater bias.

In the next section, to provide a basis for developing the questions studied here, rater cognition research and its various ramifications are discussed in more detail. Subsequently, the focus is directed to rater bias studies employing a measurement approach to the analysis of rater behavior. Then the specific questions that motivated the present research are presented.

RATER COGNITION RESEARCH

Starting around the early 1990s, an increasing number of studies have addressed the judgmental and decision-making processes involved in arriving at a particular rating. These studies are important because they set the stage for asking how these processes relate to evaluating examinee performance within the context of operational scoring sessions. The focus of the present research was on the rater cognition–behavior link regarding *writing* performance assessment. Therefore, studies of the rating processes involved in speaking performance assessment are only mentioned casually (for reviews of relevant research on speaking assessment, see A. Brown, Iwashita, & McNamara, 2005; Eckes, 2009b, 2010).

Mostly employing a qualitative approach, in particular verbal protocol analysis, researchers identified a large and varied set of decision-making strategies or reading styles that raters utilize when evaluating essays (e.g., Barkaoui, 2010; Crisp, 2008; Cumming, 1990; Cumming, Kantor, & Powers, 2002; Huot, 1993; Lumley, 2005; Milanovic, Saville, & Shuhong, 1996; Sakyi, 2000; Vaughan, 1991). For example, Vaughan (1991) identified a number of reading styles such as the "first-impression-dominates style" or the "grammar-oriented style" and concluded that similar training notwithstanding, raters may tend to differentially focus on particular performance features (e.g., content, punctuation, morphology).

Building strongly on this line of research, Wolfe (1997) advanced a rater cognition model consisting of two separate, yet interrelated components. The first component, the *framework of scoring*, comprises the mental processes through which "a text image is created, compared to the scoring criteria, and used as the basis for generating a scoring decision" (p. 89). In keeping with Freedman and Calfee (1983), Wolfe construed the text image as a mental representation of an essay that is built up while reading the essay being evaluated. The second component, the *framework of writing*, is "a mental representation of the criteria contained in the scoring rubric" (Wolfe, 1997, p. 89). According to Wolfe, both kinds of mental representation may differ from one rater to another. That is, raters may differ in the text images they create, as well as in their views of the characteristics that describe writing proficiency at different levels.

In subsequent research addressing the rater cognition model, Wolfe, Kao, and Ranney (1998) introduced the concept of *scoring focus* as a key component of the framework of writing. This concept refers to the mental weighting scheme that is implicated when raters interpret and apply

scoring criteria. Wolfe et al. assumed that raters adopt different scoring foci and that the scoring focus adopted by a particular rater may differ from the set of criteria contained in the scoring rubric. Analyzing think-aloud protocols, Wolfe et al. provided evidence that the adoption of a particular scoring focus at least in part hinges on the level of scoring proficiency a rater has achieved. Highly proficient raters tended to use a top-down approach to essay scoring, focusing on performance features that are more general. Less proficient raters tended to use a bottom-up approach to essay scoring, focusing on performance features that are more specific. In addition, less proficient raters seemed to be more likely to resort to performance features not captured in the scoring rubric.

Bejar et al. (2006) proposed a conceptually related cognitive modeling approach. Using a small sample of three teachers with varying degrees of teaching experience (5, 25, or 30 years of teaching, respectively), the authors first elicited from each rater a set of features he or she deemed important in differentiating among a number of written performances. Then, raters provided holistic scores for each performance. Based on a nonparametric regression technique, which modeled the scores awarded by each rater as a function of the set of performance features established in the previous step, Bejar et al. were able to show that raters approached the task of evaluating written performances from widely differing perspectives on the relative importance of features. Yet there was also evidence of some commonality between evaluations, at least as far as raters having similar teaching experience were concerned. That is, the features that each of the two more experienced raters deemed important were highly predictive of the other rater's score.

Discussing the nature of variability among raters, McNamara (1996) pointed out that the pattern of observed differences between raters "may lead us to think in terms of *rater types* [emphasis added]" (p. 124). Following this line of reasoning, Eckes (2008b) probed more deeply into the differences and commonalities between raters' scoring perspectives. Specifically, he advanced the hypothesis that experienced raters, that is, trained and certified raters with many years of rating experience, fall into types or classes that are clearly distinguishable from one another with respect to the importance they attach to scoring criteria. In that study, a sample of 64 raters actively involved in scoring examinee performance on a large-scale writing assessment were asked to indicate on a 4-point scale how much importance they would, in general, attach to criteria that had been in use for years. These criteria covered various performance aspects, such as fluency, completeness, and grammatical correctness.

An analysis of the criterion importance ratings yielded a joint classification of raters and criteria into six rater types, where each type was characterized by a distinct scoring profile. That is, each type exhibited a unique profile of perceived scoring criterion importance. Four of these types were defined by criteria standing out as extremely important: the *Syntax Type*, attaching high value to texts exhibiting a wide range of cohesive elements and syntactic structures; the *Correctness Type*, attaching high value to texts exhibiting only few syntactic, lexical, or orthographic errors; the *Structure Type*, attaching high value to texts that are structured according to academic requirements; and the *Fluency Type*, attaching high value to texts that can be read fluently. The remaining two types were defined by criteria to which raters gave specifically less weight: the *Nonfluency Type* and the *Nonargumentation Type*. In addition, some rater type profiles were in sharp contrast with one another, which means that raters of different types showed markedly different scoring foci. Much the same picture of distinct rater perspectives on scoring criteria emerged from a similar analysis of rater types in speaking performance assessment (Eckes, 2009b).

Note that the rater type study (Eckes, 2008b) did not aim to examine bias effects within the context of operational performance assessment. Thus, raters were explicitly asked *not* to think of a particular writing performance when providing their criterion importance ratings. In other words, the analysis was based solely on self-report data, lacking any direct connection to operational scoring behavior.

Taken together, rater cognition studies have provided ample evidence that raters perceive criteria in systematically different ways. These rater differences may add variability to test results and complicate the interpretation of scores, in that different scores may reflect slightly different versions of the overall construct. More direct evidence of rater variability in criterion perception has been provided by rater bias studies discussed next.

RATER BIAS STUDIES

Rater bias refers to a systematic pattern of rater behavior that manifests itself in unusually severe (or lenient) ratings associated with a particular aspect of the assessment situation. For example, raters may show unexpectedly high degrees of severity when scoring performance of a particular group of examinees, when scoring examinee performance on a particular task, or when using a particular scoring criterion. When raters show evidence of exercising this kind of differential severity (or leniency), they are said to exhibit *differential rater functioning* (e.g., Du, Wright, & Brown, 1996; Engelhard, 2007). The analysis of differential rater functioning is called *bias analysis* (Bachman, 2004; McNamara, 1996). Commonly, bias analyses have been performed building on a many-facet Rasch measurement (MFRM) approach (Linacre, 1989).

Several rater bias studies have addressed the differential impact of rater background variables, in particular, rater language background on rater severity (Caban, 2003; Johnson & Lim, 2009; Kim, 2009), whereas others have examined possible effects of rater training on levels of inter-rater agreement and rater severity (Elder, Barkhuizen, Knoch, & von Randow, 2007; Elder, Knoch, Barkhuizen, & von Randow, 2005; Knoch, Read, & von Randow, 2007; O'Sullivan & Rignall, 2007). As noted previously, rater bias can be studied in relation to various aspects of the assessment situation. Given the prominent role scoring criteria play in the complex process of assessing examinee performance, bias analyses examining interactions between raters and criteria have been of particular interest (e.g., Knoch et al., 2007; McNamara, 1996; Schaefer, 2008; Wigglesworth, 1993). For example, Wigglesworth (1993) found that some raters scored consistently more harshly on a criterion referring to *grammar*, whereas others scored on this criterion more leniently. She observed similar patterns of between-rater severity differences in relation to criteria such as *fluency* and *vocabulary*. The occurrence of criterion-related rater bias was specifically studied by Knoch et al. (2007). Focusing on a comparison between online and face-to-face training procedures, the researchers found that in both training groups only a few raters exhibited less bias after training, whereas others even developed new biases.

Studying rater behavior within the context of the Occupational English Test, McNamara (1996) noted a contrast between the underlying communicative orientation of the test, which downplayed the importance of performance features referring to grammatical accuracy, and the analysis of ratings (MFRM analysis, regression analysis), which revealed that raters' perceptions of *grammar* had a predominant influence on awarding test scores. This contrast remained

even though raters were trained in the communicative spirit of the Occupational English Test, indicating the presence of a specific *grammar*-related bias that was highly resistant to change.

In a study that bore a striking resemblance to the notion of rater types, Schaefer (2008) explored bias patterns of relatively inexperienced native English-speaker raters evaluating EFL essays written by Japanese university students. Raters used an analytic scoring rubric comprising six criteria: *content*, *organization*, *style and quality of expression*, *language use*, *mechanics*, and *fluency*. The MFRM Rater × Criterion interaction analysis yielded a substantial proportion of significant bias terms. By sorting the flagged interactions into unexpectedly severe or unexpectedly lenient ratings, Schaefer was able to identify subgroups of raters sharing a particular pattern of rater bias. For example, there was a subgroup of raters who exhibited unusually high severity toward *content* and/or *organization* and at the same time exhibited unusually high leniency toward *language use* and/or *mechanics*. Another subgroup showed exactly the reverse bias pattern: high leniency toward *content* and/or *organization* coupled with high severity toward *language use* and/or *mechanics*.

Schaefer's (2008) study yielded valuable insight into specific patterns of bias shared by subgroups of raters. However, that study failed to provide an explanation as to *why* some raters showed evidence of exercising differential severity/leniency toward particular criteria, whereas other raters showed exactly the reverse bias pattern. Viewed from a rater cognition perspective, the emergence of distinct rater subgroups in Schaefer's research suggests the existence of a link between criterion-related rater bias and the perceived importance of scoring criteria. In particular, one possibility might be that Schaefer's subgroup providing severe ratings on *content* and/or *organization*, and lenient ratings on *language use* and/or *mechanics*, comprised raters who perceived the first set of criteria as highly important and the second set of criteria as much less important. Systematic differences in perceived criterion importance may thus have contributed to the occurrence of group-specific bias patterns.

The purpose of the present research was to overcome the limitations of both Eckes's (2008b) rater cognition study, which failed to address operational rater bias, and Schaefer's (2008) rater bias study, which failed to account for the occurrence of differential severity/leniency toward particular criteria. Basically, this research addressed the question of how rater cognition relates to rater behavior. In so doing, the classificatory approach adopted in previous rater type studies (Eckes, 2008b, 2009b) was extended to the realm of operational rater behavior. The basic hypothesis was that rater types identified in the rater cognition study would be linked to patterns of rater bias toward scoring criteria in a large-scale writing assessment.

RESEARCH QUESTIONS

The findings from the rater cognition and rater bias studies reviewed in the previous sections suggest three conclusions. First, raters are generally subject to criterion-related bias. Second, biased use of scoring criteria is related to how raters perceive and interpret the criteria. Third, rater bias is difficult to combat through rater training. What remains currently unknown is the precise nature of the link between criterion perception and criterion-related rater bias. The rater type perspective adopted in the present research aimed to provide a basis for answering this issue.

An important step of this research was to identify rater types within the context of operational scoring sessions. In the present study, these rater types are called *operational rater types* (ORTs),

for ease of distinction from those types identified in the previous rater cognition study (Eckes, 2008b), which are called *cognitive rater types* (CRTs). Whereas CRTs show distinct patterns of attaching importance to scoring criteria, ORTs show distinct patterns of criterion-related bias exhibited in operational rating behavior. That is, it is hypothesized that raters belonging to the same ORT would show similar bias patterns (i.e., similar patterns of differential severity/leniency toward scoring criteria), whereas raters belonging to different ORTs would show dissimilar bias patterns.

The research question then becomes whether there is a link between the CRTs and the ORTs, and what the precise nature of that link is. Basically, the hypothesis advanced here draws on the notion of motivated perception, judgment, and decision making (e.g., Kunda, 1990; Weber & Johnson, 2009; Zeelenberg, Nelissen, & Pieters, 2008). According to this notion, individuals take advantage of ambiguities in stimulus properties to see objects and people in relation to strongly held beliefs or expectations. Applied to the present research, the nature of the rater cognition–rater behavior link is hypothesized to rest on the differentially perceived importance of scoring criteria. That is, to the extent that criteria are perceived to differ in importance, raters will tend to be biased toward these criteria in operational scoring sessions.[1]

More specifically, as compared to criteria perceived as relatively unimportant, examinee performance that relates to criteria perceived as relatively important will be attended to more closely and evaluated more rigorously. In effect, performance features that do not live up to expectations based on important criteria will be assigned more severe ratings. Consider a criterion that is perceived as highly important by a particular rater. This criterion is likely to be associated with strong beliefs held by that rater as to what a proficient examinee should be able to do. Then, examinee performance deviating from these beliefs will be rated more harshly, as compared to criteria perceived as less important. Moreover, the *opposite* effect may show up with criteria perceived as less important: Features of examinee performance that do not satisfy expectations based on these criteria will be rated more leniently. Put another way, to the extent that a particular rater views operationally used criteria as differing in importance, he or she will tend to score harshly on one subset of criteria (i.e., those viewed as more important) and leniently on another subset (i.e., those viewed as less important). Such a dual severity/leniency effect would indicate the occurrence of a *compensatory* rater bias, that is, on average, a given rater would not appear as being biased with respect to his or her use of scoring criteria.

To sum up, the present research addressed the following questions:

1. Can ORTs be identified with respect to a live scoring session?
2. How well are the ORTs differentiated from one another with respect to rater bias patterns?
3. Are CRTs identified in previous research linked to the ORTs?

Specifically:

3a. Are criteria perceived as highly important more closely associated with severe ratings, and
3b. Are criteria perceived as less important more closely associated with lenient ratings?

[1] It should be noted, however, that systematic differences in raters' criterion perception are just one of a number of factors possibly contributing to Rater × Criterion interactions. Other likely factors include the specific design of the scoring rubric, the time of rating, and the positioning of criteria.

To the extent that the CRT–ORT, or cognition–bias, link hypothesis is borne out in the data, an important piece of a jigsaw puzzle would be added to the emerging picture of factors accounting for the occurrence of rater variability, in particular rater bias, in operational scoring sessions. Evidence in favor of this hypothesis would also inform measures that aim at reducing rater bias through specific procedures of rater training and rater monitoring.

METHOD

Participants

In the rater cognition study (Eckes, 2008b), 64 raters participated and were classified into CRTs. Of those, only 18 raters (16 women, two men) had also participated in an operational scoring session from which the data for the present study were obtained. This session took place 3 to 4 months after the rater cognition data had been gathered. Raters were all experienced teachers and specialists in the field of German as a foreign language; that is, they all had several years of work experience, mostly 10 or more years, as language teachers and examiners at university language centers or at institutions closely affiliated with universities and were systematically trained and monitored to comply with scoring guidelines (e.g., Eckes, 2008a). Raters' ages ranged from 33 to 70 years ($M = 45.39$, $SD = 11.82$). For another group of 15 raters participating in the operational scoring session no data on perceived criterion importance were available. This group was not considered further in the examination of the cognition–bias link hypothesis but was combined with the first group of 18 raters in order to provide a more inclusive and reliable data base for estimating bias parameters in the MFRM analysis (resulting in a total of 33 operational raters).

The first group of raters, that is, those 18 raters for whom both cognition and operational rating data were available, belonged to one of three CRTs from the set of six CRTs identified previously (Eckes, 2008b). Table 1 lists the respective CRTs (i.e., types A, C, and D), as well as the number of raters per CRT that served as operational raters in the writing assessment and, thus, became part of the rater bias study.

TABLE 1
Number of Raters per Cognitive Rater Type Included in the Bias Analysis of the Operational Writing Assessment

CRT	*Label*	*No. of Raters per CRT*	*No. of Raters in Bias Analysis*
A	Syntax	17	9
C	Structure	12	5
D	Fluency	23	4

Note. Cognitive rater types (CRTs) A, C, and D refer to the importance perception clustering solution comprising six rater types (labeled A–F; Eckes, 2008b). CRT A contained raters with a focus on criteria relating to *linguistic realization* (e.g., *syntax*) and *task realization* (e.g., *completeness*). CRT C contained raters with a focus on criteria relating to *overall impression* (e.g., *structure*) and *task realization* (e.g., *argumentation*). CRT D contained raters with a focus on criteria relating to *overall impression* (e.g., *fluency*).

Instruments and Procedure

The data came from a live *Test Deutsch als Fremdsprache* (Test of German as a Foreign Language; TestDaF) examination. The TestDaF is a high-stakes test designed for foreign students applying for entry to an institution of higher education in Germany. This test measures the four language skills (reading, listening, writing, and speaking) in separate sections. Specifically, the writing section is designed to assess the examinees' ability to produce a coherent and well-structured text on a given topic taken from the academic context (for more detail, see Eckes, 2008a; Eckes et al., 2005; see also http://www.testdaf.de).

As mentioned before, in the scoring session a total of 33 raters scored the writing performance of 2,097 examinees (1,237 women, 860 men). All examinees had taken the same writing test. Each essay was scored by a single rater. In addition to their normal workload, all raters scored the same set of three carefully selected essays taken from a trialing of the writing task used operationally. These essays covered the range of relevant proficiency levels addressed in the assessment. The additional ratings served to ensure connectedness of the resulting data set such that the severity and bias measures constructed for the raters could be directly compared to one another (see Eckes, 2008a, 2011; Myford & Wolfe, 2000). Twenty-nine raters scored between 62 and 69 essays, and the remaining four raters scored 20, 26, 123, or 186 essays, respectively.

Essay ratings were carried out according to a detailed catalogue of performance aspects comprising nine criteria. The first three criteria were subsumed under the heading of *overall impression* and were more holistic in nature:

- *Fluency*: the degree to which the text can be read fluently (i.e., without impairment of understanding upon first reading the text)
- *Train of thought*: the degree to which the train of thought is consistent and coherent
- *Structure*: the degree to which the text is adequately structured (i.e., structured in accord with academic requirements)

The next three criteria referred to aspects of *task realization* and were more of an analytic kind:

- *Completeness*: the degree to which all of the points specified in the task description are dealt with
- *Description*: the degree to which important information contained in the prompt, such as a table or diagram, is succinctly summarized
- *Argumentation*: the degree to which points of view/personal considerations are recognizable (e.g., weighing the pros and cons of a particular stance)

Finally, the following (analytic) criteria referred to various aspects of *linguistic realization*:

- *Syntax*: the degree to which the text exhibits a range of cohesive elements and syntactic structures
- *Vocabulary*: the degree to which the vocabulary is varied and precise
- *Correctness*: the degree to which the text is free from morphosyntactic, lexical, or orthographical errors.

On each criterion, examinee performance was scored using a four-category scale, with scale categories designated by TestDaF levels (or TDNs, for short): *below TDN 3*, *TDN 3*, *TDN 4*, and *TDN 5*. Proficiency levels TDN 3 to TDN 5 correspond to CEFR levels B2.1 to C1.2 (Council

of Europe, 2001; see also Eckes, 2005, 2008a; Kecker & Eckes, 2010). That is, the test measures German language ability at an intermediate to high level. There is no differentiation among lower ability levels. It is just noted that TDN 3 has not yet been achieved (*below TDN 3*).

TDN levels 3 to 5 may be illustrated in terms of what an examinee can do. For example:

- *TDN 3*: The examinee can write generally comprehensible and structured texts in common study-related situations; he or she can write simply structured texts in a general academic context; linguistic and structural deficiencies may impair understanding.
- *TDN 4*: The examinee can write structured and cohesive texts in a style generally appropriate to the context in common study-related situations and in a general academic context; occasional linguistic deficiencies do not impair understanding.
- *TDN 5*: The examinee can write well-structured and cohesive texts in a style appropriate to the context using differentiated vocabulary in common study-related situations (e.g., a report for a grant awarding body) and in a general academic context (e.g., course notes, synopsis of a lecture).

Data Analysis

Many-facet Rasch analysis. The operational rating data were analyzed by means of the computer program FACETS (version 3.68; Linacre, 2011). The program used the ratings that raters awarded to examinees to estimate individual examinee proficiencies, rater severities, criterion difficulties, and scale category difficulties. Building on findings from earlier MFRM analyses of writing performance assessments (e.g., Eckes, 2004, 2005), the Rasch model used here was a three-facet rating scale model, that is, all ratings were assumed to follow the same scale structure (see also Eckes, 2009a, 2011; Myford & Wolfe, 2003, 2004). The facets considered in the MFRM analysis were examinees, raters, and criteria.

Based on the parameter estimates obtained for examinees, raters, and criteria, various interactions involving the rater facet were examined. As mentioned previously, the focus was on interactions between raters and criteria. That is, for each and every rater–criterion combination, the criterion-related interaction (or bias) analysis yielded an estimate of the extent to which a rating provided by a particular rater on a particular criterion was higher (or lower) than expected, given the rater's measure of severity and the criterion's measure of difficulty.

These estimates are called *bias measures* (reported in logits). Bias measures greater than 0 logits indicate that observed scores were higher than expected based on the model; that is, raters were more lenient on the criteria involved than on other criteria. Conversely, bias measures less than 0 logits indicate that observed scores were lower than expected; that is, raters were more severe on the criteria involved than on other criteria. Bias measures provided the input to the cluster analysis discussed next. Unlike the commonly reported standardized fit statistics (or *t* statistics), which test the null hypothesis that the data fit the model "perfectly," bias measures indicate whether the data fit the model "usefully" (Linacre, 2003). For the purposes of cluster interpretation, an (absolute) bias measure of at least 0.50 logits was taken as a lower limit for a tendency to score more severely (or more leniently) than expected. Within the context of item bias or DIF analysis, this so-called half-logit rule (Draba, 1977) has been shown to yield reliable results (e.g., Elder, McNamara, & Congdon, 2003; Tristán, 2006; Wang, 2000).

Cluster analysis of bias measures. To identify ORTs, the bias measures resulting from the Rater × Criterion interaction analysis were analyzed by means of a hierarchical clustering algorithm generally deemed appropriate for the analysis of interval-scale data (i.e., Ward's method, also known as the incremental sum of squares algorithm; Ward, 1963; see Everitt, Landau, Leese, & Stahl, 2011; Gordon, 1999). This algorithm was run using the computer program ClustanGraphics (version 8.06; Wishart, 2006). Ward's method yields a series of monotonically increasing fusion values, each of which is the smallest increase in the total within-cluster error sum of squares resulting from the fusion of two clusters at the preceding level of the clustering hierarchy. Each cluster in the hierarchy represents a subgroup of raters sharing a particular pattern of providing severe (and/or lenient) ratings on some criteria but not on others.

To guard against interpreting clusters that do not adequately recover the structure inherent in the bias measures and to decide on the appropriate number of clusters, that is, on the appropriate number of ORTs, the following approach was adopted. First, in addition to Ward's method two other hierarchical algorithms were employed (i.e., Complete Linkage, Average Linkage; see Everitt et al., 2011; Gordon, 1999). These algorithms yielded highly similar clustering results, attesting to the stability of the obtained solution. Therefore, only the ORTs as identified by Ward's method are discussed next. Second, a bootstrap cluster validation procedure was applied to the resulting clustering hierarchy. In this procedure, a clustering hierarchy obtained for a given data set is compared with a series of clustering hierarchies generated at random from the same data set. The best clustering solution, that is, the most valid hierarchy level, is taken to be the one that exhibits the greatest departure from randomness (Wishart, 2005).

RESULTS

Many-Facet Rasch Analysis

Interrater agreement. Use of an MFRM modeling approach implies the assumption that raters acted as independent experts (e.g., Eckes, 2011; Linacre, 1997). To check this assumption, interrater agreement statistics were computed. The observed proportion of exact agreements between raters (building on the common ratings of preselected essays) was 42.1%; based on MFRM model parameter estimates the expected proportion of exact agreements was 41.1%. Inserting these proportions into the Rasch-kappa index (Eckes, 2011; Linacre, 2011) yielded a value of 0.02, which was reasonably close to zero, the value that would be obtained under Rasch-model conditions. This suggested that raters were able to rate the common essays based on the same general point of view in terms of their understanding of the criteria, yet at the same time, raters were not overly dependent on one another. From a more traditional perspective, interrater reliability was assessed by means of Cronbach's alpha (computed across criteria for each essay separately). The values ranged from .86 to .91, attesting to a sufficiently high agreement between raters.

Rater measurement results. To provide a basis for the presentation of the bias and cluster analysis results, the rater calibrations are briefly discussed first (see Appendix A for a graphical illustration of the complete set of calibrations).

The variability across raters in their level of severity was substantial. The rater severity measures showed a 4.06-logit spread, which was more than one fourth of the logit spread observed for examinee proficiency measures.[2] The most severe rater (i.e., Rater 45A) was from CRT A, as was the most lenient rater (i.e., Rater 35A). Clearly, the separation statistics (e.g., Eckes, 2011; Myford & Wolfe, 2003) confirmed that rater severity measures were far from being homogeneous: The rater homogeneity index, Q, computed as a fixed chi-square was highly significant, $Q(32) = 3{,}529.9$, $p < .001$, indicating that at least two raters did not share the same parameter (after allowing for measurement error); furthermore, the number of strata index, H, showed that within the present group of raters there were about 14 statistically distinct strata of severity ($H = 13.83$), and the reliability of rater separation, R, was close to its theoretical maximum ($R = .99$).

Another issue concerned the degree to which each rater made consistent use of the scale categories across examinees and criteria. Within-rater consistency was assessed by means of mean-square fit statistics, which have an expected value of 1 and can range from 0 to infinity (e.g., Eckes, 2011; Engelhard, 2002; Myford & Wolfe, 2003). In the present analysis, most of the 33 raters had fit statistics that stayed within a narrowly defined fit range (lower-control limit = 0.80, upper-control limit = 1.20; e.g., Bond & Fox, 2007; Eckes, 2011); that is, these raters performed in a highly consistent manner. Thirteen raters had fit statistics outside that narrow fit range, of which five raters exhibited a tendency to show moderate misfit (i.e., infit values between 1.25 and 1.30) and another five raters exhibited a tendency to show moderate overfit (i.e., infit values between 0.70 and 0.78). The remaining three raters showed somewhat more misfit (i.e., infit values between 1.32 and 1.41). Yet none of the raters could be said to provide ratings that were overly inconsistent.

Rater × Criterion bias analysis. To investigate whether each rater maintained a uniform level of severity across criteria, or whether particular raters scored on some criteria more harshly/leniently than expected on the basis of the many-facet Rasch model, a two-way interaction (i.e., Rater × Criterion) analysis was performed. In addition, a Rater × Examinee analysis and a three-way Rater × Examinee × Criterion analysis were also run.

As shown in Table 2, the percentage of (absolute) t values equal or greater than 2 for the Rater × Examinee and Rater × Examinee × Criterion interactions were generally fairly low. Quite in contrast to this was the high percentage of Rater × Criterion interactions that were associated with substantial differences between observed and expected ratings; all of these differences were statistically significant at the .05 level. Obviously then, raters experienced considerable difficulty in applying the criteria in a manner that stayed sufficiently close to their overall level of severity or leniency. The cluster analysis of the Rater × Criterion bias measures reported next was to shed light on the nature of these bias patterns.

Cluster Analysis of Bias Measures

For the purposes of comparison, Appendix B presents the profiles of perceived criterion importance that characterized CRTs A, C, and D, as identified in the Eckes (2008b) rater cognition

[2] The distribution of examinee proficiency measures pointed to pronounced between-examinee differences in writing proficiency. This was not unexpected given the typical composition of the TestDaF candidature, ranging from beginners just trying their luck at the TestDaF to highly advanced learners of German coming close to native-language proficiency levels.

TABLE 2
Summary Statistics for the Many-Facet Rasch Interaction Analysis

Statistics	*Rater × Examinee*	*Rater × Criterion*	*Rater × Examinee × Criterion*
N combinations	2,106	297	18,950
% large *t* values[a]	0.8	44.4	1.5
% sign. *t* values[b]	0.5	44.4	0.0
Min-*t* (*df*)	−3.57* (8)	−7.89* (62)	−3.45 (1)
Max-*t* (*df*)	3.66* (8)	8.96* (181)	3.43 (1)
M	0.00	0.00	−0.03
SD	0.32	2.64	0.79

[a]Percentage of absolute *t* values (standardized bias measures) ≥ 2. [b]Percentage of *t* values statistically significant at $p < .05$.
*$p < .01$.

study. That is, Table B1 shows which criteria raters of a given cognitive type considered highly important, denoted by a plus sign, or less important, denoted by a minus sign; an empty cell signifies that the respective criterion was viewed as moderately important. Thus, for example, CRT A, or Syntax Type raters, viewed as highly important *train of thought*, *completeness*, *argumentation*, *syntax*, and *vocabulary*; as moderately important *fluency*, *description*, and *correctness*; and as less important *structure*. As can be seen, Structure Type and Fluency Type raters each had a profile clearly distinguished from that of Syntax Type raters. The only criterion viewed as highly important by raters of all three types was *train of thought*.

The hierarchical clustering analysis of the bias measures (in logits) yielded the tree diagram shown in Figure 1. Using the bootstrap validation procedure discussed above, four clusters of operational raters were identified. Each of these clusters represented an operational rater type, labeled ORT 1 to ORT 4.

On the left-hand side of the tree diagram, raters are designated by their identification number as used in the rater cognition study (Eckes, 2008b) along with the corresponding CRT identifier (A, C, or D). For example, the label "02C" refers to Rater 02 shown in the cognition study to belong to Rater Type C (Structure Type). The clustering algorithm placed raters in a stepwise manner into more and more inclusive classes according to the congruence in criterion-related bias measures. In the diagram, the height of the point (from left to right) where two raters were first placed in the same class represents the distance between these raters in terms of their bias tendencies. These distances are shown along the scale at the bottom of Figure 1, labeled "Increase in Sum of Squares." The larger the distance between raters is, the lower the congruence in these raters' criterion-related biases. Thus, minimum distance, or highest congruence in bias measures, was obtained for Raters 11A and 31C from ORT 3. Conversely, maximum distance was present between raters belonging to ORT 1 (or ORT 2), on one hand, and raters belonging to ORT 3 (or ORT 4) on the other.

As can readily be seen, ORT 1 comprised four raters, three of these belonging to CRT A. These raters were joined by a rater from a different CRT (i.e., Rater 02C). Note that in the rater cognition study (Eckes, 2008b), CRTs A and C were more closely aligned to each other than each of these to CRT D; that is, CRT D raters exhibited a view of criterion importance widely different from that of CRT A or CRT C raters. Therefore, CRT A raters grouped together with CRT C

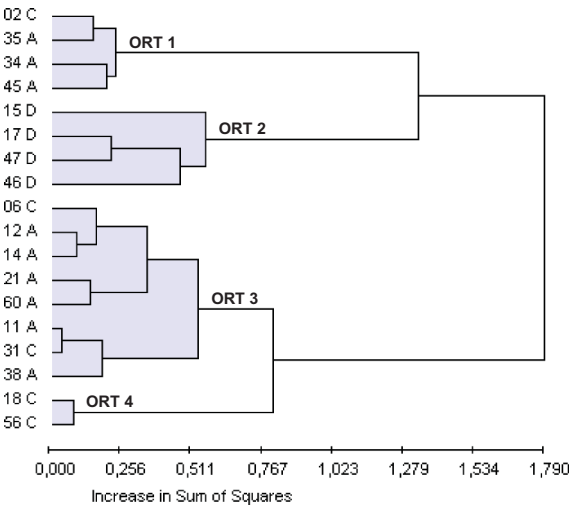

FIGURE 1 Hierarchical clustering solution for raters based on criterion-related bias measures. *Note.* Raters are designated by their identification number from the rater cognition study (Eckes, 2008b) along with the corresponding CRT identifier (A, C, or D). Each rater belongs to one of four operational rater types (ORT 1 to ORT 4) (color figure available online).

raters in the same ORT would be much less of a departure from the hypothesized link than CRT A (or CRT C) raters grouped together with CRT D raters. In fact, as Figure 1 shows, there was one other ORT (i.e., ORT 3) that contained a mixture of raters from CRT A and CRT C, and the remaining two ORTs were exclusively matched with raters from CRT D (i.e., ORT 2) or from CRT C (i.e., ORT 4), respectively.

Next, patterns of bias toward specific scoring criteria were compared between the ORTs. Figure 2 displays the bias profiles for the set of four ORTs. Each profile shows the mean bias measures obtained for a particular ORT. Note first that the profiles differed sharply from each other. The most distinctive profile was obtained for ORT 2. As discussed earlier, ORT 2 was composed solely of CRT D raters. Referring back to Appendix B, these raters differentially perceived *fluency* and *train of thought* as highly important. The profile depicted in Figure 2 confirmed that raters belonging to ORT 2 scored particularly harshly on these two criteria, as indicated by the negative bias measures; all other criteria were associated with positive mean bias measures, that is, ORT 2 raters tended to score more leniently on criteria such as *completeness*, *description*, and *syntax*. Most of these criteria were perceived as less important by this particular type of raters. The only exception was *syntax*, a criterion on which ORT 2 raters scored very leniently though it was not included in the list of criteria perceived as less important by CRT D raters.

Table 3 presents differential severity/leniency biases observed for raters from ORT 1 to ORT 4 combined with criterion importance ratings provided by raters belonging to CRT A, C, or D.

Perceived importance of a given criterion is indicated by a plus sign (high importance) or minus sign (low importance). Bias direction is indicated by capital letter L (leniency bias) or S (severity bias). Hence, a combined symbol such as "+/S" refers to a criterion that is perceived as highly important *and* is associated with a severity bias. As mentioned previously, the half-logit

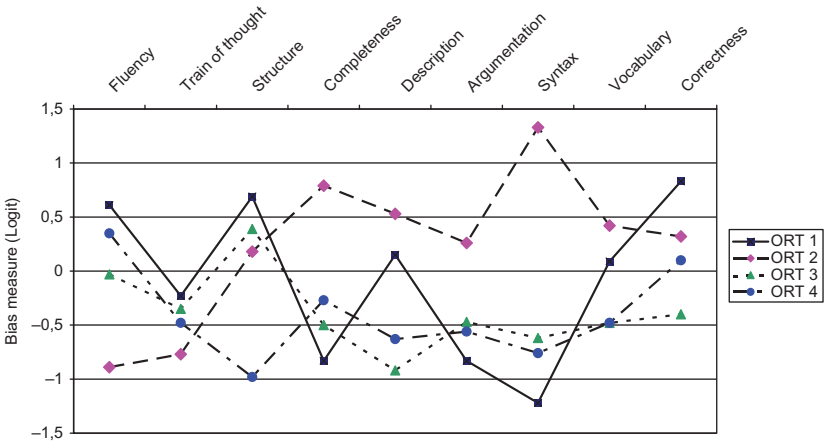

FIGURE 2 Bias diagram for operational rater types ORT 1 to ORT 4. *Note.* Mean bias measures for each ORT are shown in logits (positive values indicate leniency bias, negative values indicate severity bias) (color figure available online).

TABLE 3
Severity and Leniency of Four Operational Rater Types in Relation to Criterion Importance Profiles of Cognitive Rater Types A, C, and D

		Overall Impression			*Task Realization*			*Linguistic Realization*		
ORT	*CRT*	*Fluency*	*Train of Thought*	*Structure*	*Completeness*	*Description*	*Argumentation*	*Syntax*	*Vocabulary*	*Correctness*
1	A/C	L	+	L	S		+/S	S		L
2	D	+/S	+/S		–/L	–/L		L	–	–
3	A/C		+		S	S	+	S	S	
4	C		+/S	+/S		+/S	+/S	S		

Note. Operational rater types (ORTs) 1 to 4 refer to the bias measures clustering solution comprising four operational rater types. Cognitive rater types (CRTs) A, C, and D refer to the importance perception clustering solution comprising six cognitive rater types (labeled A–F; Eckes, 2008b). + = high importance; – = low importance; L = leniency bias (i.e., average bias measure ≥ 0.5 logits); S = severity bias (i.e., average bias measure ≤ –0.5 logits).

rule (Draba, 1977) was applied to detect criteria that were rated more severely (or more leniently) than expected. Note also that ORTs 1 and 3 were composed of raters from two different CRTs (i.e., CRT A and CRT C); these two ORTs will be referred to as *mixed* rater types. Specifically, CRT A and CRT C raters differed widely in their view of *structure* and, somewhat less, in their views of four other criteria (*completeness*, *description*, *syntax*, and *vocabulary*), but concurred in their views of *train of thought* and *argumentation* (see Appendix B). To arrive at cluster characterizations that focused on bias tendencies common to raters of a given mixed type, only the importance marker for these two agreed-upon criteria was entered in Table 3.

Looking first at mixed rater type ORT 1, there was one criterion, *argumentation*, in which the hypothesized link was borne out. ORT 1 raters viewed this criterion as highly important and were

differentially severe when using it operationally. Two other criteria were associated with a severity bias (*completeness*, *syntax*) but lacked an indication as to perceived importance. Conversely, *train of thought* had a high importance rating but lacked a severity bias. Finally, three criteria (*fluency*, *structure*, *correctness*) showed a leniency bias but were not specifically associated with a low importance rating.

As mentioned previously, raters from the "pure" ORT 2 (i.e., all ORT 2 raters also belonged to CRT D) exhibited the clearest evidence of the hypothesized cognition–bias link: *fluency* and *train of thought* were perceived as highly important, and both were used in a severe manner operationally. The reverse pattern showed up for *completeness* and *description* (i.e., low perceived importance combined with a leniency bias). However, criteria referring to *linguistic realization* failed to show a consistent link between cognition and behavior. Looking only at the occurrence of a severity and/or leniency bias, ORT 2 raters seemed to exhibit a kind of compensatory bias similar to the bias pattern reported by Schaefer (2008). Thus, a tendency to score harshly on criteria referring to *overall impression* was in a sense compensated by a tendency to score leniently on criteria referring to *task realization*. A similar picture emerged for ORT 1 raters, with an additional compensatory relation between criteria referring to *linguistic realization* (i.e., *syntax*, *correctness*).

ORT 3 raters, representing another "mixed" type, exhibited a severity bias in four cases with no indication of any specific relation to perceived criterion importance. Quite in contrast, ORT 4 raters, representing another "pure" type, exhibited severity bias on five criteria, four of which were perceived as highly important. Different from ORT 1 and ORT 2, raters belonging to ORT 3 or ORT 4 did not show any signs of compensatory biases.

Considering the range of bias measures per criterion (see Figure 2), *syntax* was associated with the largest difference between ORTs, followed by *structure*, *completeness*, and *fluency*. The smallest range of bias measures (i.e., 0.54 logits) was observed for *train of thought*; that is, this criterion differentiated only weakly between the ORTs in terms of bias size. Note also that *train of thought* was the only criterion that raters from all three CRTs perceived as highly important.

SUMMARY AND DISCUSSION

The present research examined the link between rater cognition and judgmental bias within the context of performance-based language assessment. In particular, the aim was to probe into the relation between criterion importance ratings and differential severity/leniency toward criteria routinely used in a large-scale writing assessment. Building on previous research into rater types (Eckes, 2008b), a distinction was made between CRTs and ORTs. CRTs were defined by distinct patterns of differentially perceived criterion importance, whereas ORTs were defined by distinct patterns of differential severity/leniency bias exhibited in a live scoring session.

The first research question asked whether ORTs could be identified within the context of a live examination of writing performance. Building on a many-facet Rasch measurement approach to construct measures of criterion-related rater bias, and using the obtained bias measures as input to a cluster-analytic procedure, raters were successfully classified into four ORTs. Cluster validation confirmed that these four rater types reliably recovered the structure inherent in the bias data.

The second research question asked whether these ORTs would be well differentiated from one another with respect to criterion-related bias patterns. That is, this part of the operational rater

type hypothesis implied that raters belonging to the same ORT would exhibit similar tendencies to score more harshly and/or more leniently on particular criteria and that these scoring tendencies would be clearly distinguishable from those tendencies exhibited by raters belonging to each of the other ORTs. Although not all criteria were subject to biases to a similar degree, a closer look at the bias profiles indeed revealed that the ORTs were well separated from one another in terms of differential severity/leniency. For example, one ORT (i.e., ORT 2) exhibited a strong severity bias when scoring on *fluency* and *train of thought*, but a strong leniency bias when scoring on *syntax*. This specific pattern did not show up for any of the other three ORTs. Quite to the contrary, ORT 1 raters tended toward severe ratings on *syntax* and to lenient ratings on *fluency*.

In sum, there was some evidence speaking in favor of first two research questions. This evidence formed the point of departure for addressing the third research question. In terms of the distinction between cognitive and operational rater types, the question was actually twofold: (a) Are the criterion importance ratings characterizing a particular CRT linked to operational rating behavior, and if so, (b) what is the precise nature of that link?

Drawing on a growing body of research into rater bias patterns, I advanced the hypothesis that criteria perceived as highly important were differentially associated with a severity bias, and criteria perceived as less important were differentially associated with a leniency bias. That is, a rater considering a particular criterion as highly important, and other criteria as less important, would tend to score on this criterion more harshly than expected based on his or her overall severity measure and the overall criterion difficulty measure. By the same token, a rater considering a particular criterion as less important, and other criteria as highly important, would tend to score on this criterion more leniently than expected based on his or her overall severity measure and the overall criterion difficulty measure.

In considering the evidence on the CRT–ORT link hypothesis, it is important to recall that two of the four ORTs were composed of raters belonging exclusively to a particular CRT identified in the previous rater cognition study (Eckes, 2008b). That is, ORT 2 was exclusively composed of CRT D raters, and ORT 4 was exclusively composed of CRT C raters. In contrast to these "pure" ORTs, two other rater types were of the "mixed" kind, being composed of raters from two different, though related, CRTs (i.e., CRT A *and* C).

For the "pure" ORTs, the hypothesis of a link between criterion importance ratings and criterion-related differential severity/leniency received fairly strong support. ORT 2 raters viewed *fluency* and *train of thought* as highly important, and these criteria were associated with a severity bias; in addition, ORT 2 raters viewed *completeness* and *description* as less important, and these criteria were associated with a leniency bias. Only *vocabulary* and *correctness* were not in line with this pattern. Similarly, raters belonging to ORT 4 viewed four criteria as highly important, and all of these were associated with a severity bias. Note that the hypothesis only states that perceiving a criterion as highly important (or less important) is associated with a severity bias or leniency bias, respectively; the hypothesis is not disconfirmed when a criterion that is viewed as moderately important is associated with a severity (or leniency bias), which happened with *syntax* for ORTs 2 and 4.

Much weaker evidence in favor of the cognition–behavior link hypothesis, though, was obtained for ORTs 1 and 3, both of which were "mixed" rater types. It should be noted, however, that among the criteria marked as + or – in Table 3, no cognition–bias pattern contrary to the hypothesis was observed; that is, in not a single case was high perceived criterion importance coupled with a leniency bias, or low perceived criterion importance coupled with a severity bias.

Taken together, the present results suggest that raters' perception of scoring criteria has an impact on the way criteria are being used in operational scoring sessions. More specifically, the results suggest an explanation why, in Schaefer's (2008) study, some criteria were rated harshly by one group of raters and leniently by another group: The mediating variable is likely to be perceived criterion importance, with highly important criteria linked to severity bias, and less important criteria linked to leniency bias.

Moreover, Schaefer observed in his study a kind of compensatory bias; that is, some raters scored harshly on one subset of criteria and leniently on another subset, and vice versa. The present findings tended to confirm Schaefer's observation for two of the four ORTs. Thus, whereas ORT 1 scored leniently on criteria referring to *overall impression* (i.e., *fluency*, *structure*) and harshly on criteria referring to *task realization* (i.e., *completeness*, *argumentation*), the reverse bias pattern was obtained for ORT 2, with harsh ratings on *overall impression* criteria (*fluency*, *train of thought*) and lenient ratings on *task realization* criteria (i.e., *completeness*, *description*). At present, it seems unlikely that these compensatory biases result from a deliberate decision-making strategy employed by raters. Yet the precise nature of these biases and the conditions facilitating their occurrence are topics worth examining in future research.

In this study, operational raters provided essay ratings up to four months after the criterion importance rating data had been gathered. This delay may have had an impact on the strength of the proposed link between rater cognition and rater behavior. In particular, raters may have changed their perspectives on criterion importance on various grounds such that their operational ratings may have been governed significantly less by their views of criterion importance declared several months earlier. This change in perspective may at least in part account for the weak evidence in favor of the cognition–bias link obtained for ORTs 1 and 3. Put differently, a more stringent test of the link hypothesis would have been to gather the criterion importance data immediately before or after the live scoring session. In addition, CRTs were derived on the basis of criterion importance ratings abstracted from the context of a specific writing task and of specific examinee performance on that task. Hence, this kind of reduced compatibility between (abstract) criterion ratings and (specific) criterion use may also have contributed to the weak cognition–bias link observed in a number of cases.

Employing a design of concurrent data collection, the identification of CRTs and ORTs may be based on better comparable rater samples. In the present investigation, only 18 of the 64 raters completing the criterion perception questionnaire served as operational raters as well. It goes without saying that a larger sample of raters providing both criterion importance and operational essay ratings would yield a more reliable data basis for studying the hypothesized cognition–behavior link. Moreover, within the context of such a design, issues of temporal stability and change of criterion perceptions and their impact on operational rating behavior may be addressed in a straightforward manner. For example, it could be asked if, and possibly in which ways, the fairly strong relation between the CRTs and ORTs 2 and 4 would change from one particular point in time to the next. Questions like these may suitably be addressed within Myford and Wolfe's (2009) MFRM framework for studying various aspects of rating behavior over time.

Since the early 1990s, rater performance has been studied from a number of different angles, including purely quantitative approaches (e.g., Engelhard, 1994; Xi & Mollaun, 2009), more qualitatively based approaches (e.g., A. Brown, 2007; Meiron & Schick, 2000), or a mix of quantitative and qualitative approaches (e.g., Kim, 2009; Weigle, 1999). Whichever approach is chosen in a particular study, it is understood that multiple research methodologies are called for to gain deeper insight into the intricate nature of rater performance and the variability in

ratings associated with that performance (Lane & Stone, 2006; Lumley & Brown, 2005). In particular, the quantitative approach adopted in the present study was limited in that it had a strong outcome orientation; that is, it addressed the variation in the final ratings awarded to examinees and did not look at raters' cognitive processes involved in arriving at a particular rating. Qualitative approaches to rater variability have the advantage of allowing researchers to examine more process-oriented issues of performance assessment (e.g., Banerjee, 2004; Lazaraton, 2008; Lumley, 2005). Specifically, within the context of the present line of research, use could be made of concurrent verbal protocols to provide data on the criteria raters attend to while scoring examinee performance.[3] This could also help to ameliorate the problem of reduced compatibility between criterion ratings and criterion use previously mentioned. More generally, future research may enrich the present classificatory approach to rater variability by probing into patterns of judgmental and decision-making processes distinctly characterizing each rater type.

CONCLUSION

In previous research, rater types have been conceptualized solely in cognitive terms, distinguishing between raters mainly on the basis of self-reported perception of criterion importance. Whether, and how, these perceptual differences were related to behavioral differences found in operational scoring sessions remained unclear. The present research aimed at bridging the gap between rater cognition and rater behavior, taking up and extending the rater type approach.

When, as demonstrated in this study, perceptions of differential criterion importance translate into differential severity/leniency bias exhibited in operational scoring sessions, rater training may benefit from taking cognitive rater types and their distinctive perceptual patterns into account. For this purpose, rater monitoring activities could be extended by gathering criterion importance ratings, on one hand, and by performing Rater × Criterion interaction analysis of operational rating behavior, on the other. Combining rater cognition data and rater bias analysis on a routine basis may serve to identify potential weaknesses in the rating procedure and options for improvement. Specifically, knowing which rater type a rater belongs to can inform efforts to change that rater's perceptual focus, that is, to achieve a more balanced perception of criterion importance and thus to reduce the probability that he or she will exhibit differential severity/leniency toward criteria in scoring essays. Another factor that may come into play is the overall proficiency level of examinees, such that the occurrence of differential severity/leniency also depends on an interaction between the particular type of rater and the level of the performance rated. Clearly, research along these lines holds a lot of promise. Yet, it is still as true as it was about 20 years ago that "much more work needs to be done on the definition of consistent rater types" (McNamara & Adams, 1991, p. 23).

REFERENCES

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Banerjee, J. (2004). Qualitative analysis methods. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching,*

[3] This idea was suggested to me by one of the reviewers.

assessment (Section D). Strasbourg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/manuel1_EN.asp?#P19_2121

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

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49–81). Mahwah, NJ: Erlbaum.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 98–139). Cambridge, UK: Cambridge University Press.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph Series, MS-29). Princeton, NJ: Educational Testing Service.

Brown, G. T. L. (2010). The validity of examination essays in higher education: Issues and responses. *Higher Education Quarterly*, *64*, 276–291.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, *21*, 1–44.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, *38*, 247–264.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*, 31–51.

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

Draba, R. E. (1977). *The identification and interpretation of item bias* (MESA Memorandum No. 25). Retrieved from http://www.rasch.org/memo25.htm

Du, Y., Wright, B. D., & Brown, W. L. (1996, April). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Conference of the American Educational Research Association, New York, NY.

Eckes, T. (2004). Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im "Test Deutsch als Fremdsprache" (TestDaF) [Rater agreement and rater severity: A many-facet Rasch analysis of performance assessments in the "Test Deutsch als Fremdsprache" (TestDaF)]. *Diagnostica*, *50*, 65–77.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*, 197–221.

Eckes, T. (2008a). Assuring the quality of TestDaF examinations: A psychometric modeling approach. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity–Proceedings of the ALTE Berlin Conference May 2005* (pp. 157–178). Cambridge, UK: Cambridge University Press.

Eckes, T. (2008b). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155–185.

Eckes, T. (2009a). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/manuel1_EN.asp?#P19_2121

Eckes, T. (2009b). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt, Germany: Lang.

Eckes, T. (2010). The TestDaF implementation of the SOPI: Design, analysis, and evaluation of a semi-direct speaking test. In L. Araújo (Ed.), *Computer-based assessment (CBA) of foreign language speaking skills* (pp. 63–83). Luxembourg: Publications Office of the European Union.

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

Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & Tsagari, C. (2005). Progress and problems in reforming public language examinations in Europe: Cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France, and Germany. *Language Testing*, *22*, 355–377.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*, 37–64.

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

Elder, C., McNamara, T., & Congdon, P. (2003). Rasch techniques for detecting bias in performance assessments: An example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied Measurement*, *4*, 181–197.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Erlbaum.

Engelhard, G. (2007). Differential rater functioning. *Rasch Measurement Transactions*, *21*, 1124.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Chichester, UK: Wiley.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). New York, NY: Longman.

Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, *41*, 337–373.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton Press.

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

Kecker, G., & Eckes, T. (2010). Putting the Manual to the test: The TestDaF–CEFR linking project. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft Manual* (pp. 50–79). Cambridge, UK: Cambridge University Press.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187–217.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26–43.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education/Praeger.

Lazaraton, A. (2008). Utilizing qualitative methods for assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (2nd ed., pp. 197–209). New York, NY: Springer.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1997). Investigating judge local independence. *Rasch Measurement Transactions*, *11*, 546–547.

Linacre, J. M. (2003). Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions*, *17*, 918.

Linacre, J. M. (2011). *A user's guide to FACETS: Rasch-model computer programs*. Chicago, IL: Winsteps.com.

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

Lumley, T., & Brown, A. (2005). Research methods in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 833–855). Mahwah, NJ: Erlbaum.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

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

McNamara, T. F., & Adams, R. J. (1991, March). *Exploring rater behaviour with Rasch techniques*. Paper presented at the 13th Annual Language Testing Research Colloquium, Princeton, NJ.

Meiron, B. E., & Schick, L. S. (2000). Ratings, raters and test performance: An exploratory study. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 153–174). Cambridge, UK: Cambridge University Press.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 92–114). Cambridge, UK: Cambridge University Press.

Myford, C. M., & Wolfe, E. W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs* (TOEFL Tech. Rep. No. TR-15). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189–227.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*, 371–389.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS Writing Module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 446–478). Cambridge, UK: Cambridge University Press.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129–152). Cambridge, UK: Cambridge University Press.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*, 465–493.

Tristán, A. (2006). An adjustment for sample size in DIF analysis. *Rasch Measurement Transactions*, *20*, 1070–1071.

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

Wang, W.-C. (2000). Modeling effects of differential item functioning in polytomous items. *Journal of Applied Measurement*, *1*, 63–82.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*, 53–85.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145–178.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*, 305–335.

Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (2nd ed., pp. 111–122). New York, NY: Springer.

Wishart, D. (2005). Number of clusters. In B. S. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, (pp. 1442–1446). New York, NY: Wiley.

Wishart, D. (2006). *ClustanGraphics primer: A guide to cluster analysis* (4th ed.). Edinburgh, Scotland: Clustan.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*, 83–106.

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

Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps?* (TOEFL iBT Research Rep. No. TOEFLiBT-11). Princeton, NJ: Educational Testing Service.

Zeelenberg, M., Nelissen, R., & Pieters, R. (2008). Emotion, motivation, and decision making: A feeling-is-for-doing approach. In H. Plessner, C. Betsch, & T. Betsch (Eds.), *Intuition in judgment and decision making* (pp. 173–189). New York, NY: Erlbaum.

APPENDIX A

Logit	Examinee	Rater	Criterion	Scale
	High	*Severe*	*Hard*	(TDN 5)
6	.			
	*.			
	.			
5	**.			
	*.			
	***.			
4	****.			
	****.			
	*****.			
3	******			-----
	*****.			
	****.			
2	*****.			
	********.	45A *		TDN 4
	*******			
1	********.			
	**********	31C 47D * * *	argumentation correctness	
	*********.	02C 15D 18C 34A 56C * * * *	description	
0	*********.	06C 11A 21A 38A *	vocabulary	-----
	*******.	12A * * * *	train of th. fluency compl. syntax	
	*******.	17D 46D *	structure	
−1	*******.			
	*******.	14A 60A *		
	******.			TDN 3
−2	******.			
	****.	35A		
	****.			
−3	***.			-----
	***.			
	**.			
−4	**.			
	*.			
	.			
−5	*.			
	.			
	.			
−6				
	.			
−7				
	.			
−8				
	Low	*Lenient*	*Easy*	(below 3)

FIGURE A1 Variable map from the many-facet Rasch analysis of the TestDaF writing assessment. *Note.* Each star in the second column represents 12 examinees, and a dot represents fewer than 12 examinees. In the third column, each star represents a rater who did not take part in the rater cognition study (Eckes, 2008b). Raters who took part in that study are designated by their identification number from that study along with the corresponding CRT identifier (A, C, or D). The horizontal dashed lines in the fifth column indicate the category threshold measures for the four-category rating scale.

APPENDIX B

TABLE B1
Criterion Importance Profiles of Cognitive Rater Types A, C, and D

		Overall Impression			*Task Realization*			*Linguistic Realization*		
CRT	*Label*	*Fluency*	*Train of thought*	*Structure*	*Complete-ness*	*Descrip-tion*	*Argumen-tation*	*Syntax*	*Vocabu-lary*	*Correct-ness*
A	Syntax		+	−	+		+	+	+	
C	Structure		+	+		+	+			
D	Fluency	+	+		−	−			−	−

Note. CRTs A, C, and D refer to the importance perception clustering solution comprising six rater types (labeled A–F; Eckes, 2008b). CRT = cognitive rater type; + = high importance; − = low importance.