Assessment-Relevant Stimuli and Judging of Writing Performances – From Micro-Judgments to Macro-Judgments

ABSTRACT

The contemporary practice of rating writing performances is grounded in an approach known as judging, which is done to avoid paying conscious attention to discrete elements in texts. Instead, it involves accounting for the overall impression made by a writing performance. However, studies have indicated that while this may be true on a conscious level, the concrete stimuli in texts still preconsciously influence the forming of such overall impressions. What is left largely unnoticed is that most assessment-relevant stimuli require the use of judging to be perceived as such. This implies that an overall, macro-judgment of a writing performance (expressed normally as a score) comprises individual (and largely preconsciously generated) micro-judgments coming together into a complex and non-linear combination-count. The paper presents an argument in favour of such a composition of judgments, demonstrates it empirically by means of a case study, and then discusses the wider consequences of this different perspective on judging.

Keywords: writing assessment, judging, macro-judgment, micro-judgment, assessment-relevant stimuli
1 Introduction: The Dual Nature of Judgments

Current practice in the assessment of writing is grounded in an approach known as judging. The process of judging is understood here as involving raters not paying conscious attention to discrete elements of writing performances, and instead having them account for the overall impression that a writing performance leaves in the form of a rating (i.e., score or grade). The study at hand aims to demonstrate how an overall impression of any writing performance being evaluated in any assessment context, representing a (macro-)judgment of it made by a human rater, comprises, at its core, a complex and non-linear combination-count of individual (micro-)judgments. Each micro-judgment is itself triggered by individual assessment-relevant text features functioning as stimuli for the rater, highlighted by the expectations of a specific assessment context1 and then being preconsciously recognized and mentally noted as such.

Starting with existing rater cognition accounts of what assessment-relevant features of this kind may be, according to raters’ own testimonies (Cai 2015; Chalhoub-Deville 1995; Cumming 1990; Cumming, Kantor, and Powers 2001; Eckes 2008; Milanovic, Saville, and Shuhong 1996; Mumford and Atay 2021; Sakyi 2000; and more), the study references one widely flagged and well-known type of individual assessment-relevant stimuli as an example, namely errors.

Because each instance of recognizing an error in a performance requires the rater to make a judgment relevant to both the operating language norm and to the specific assessment context as the pertinent sounding boards, the study argues that the overall macro-judgment of a writing performance is, at least partly, derivative from individually incurred micro-judgments triggered by raters recognizing errors as one set of influential stimuli. Other assessment-relevant stimuli (such as “assessment-positive” features, see Dobrić (2023)) can be assumed as being treated in a similar manner by raters.

Indeed, the earliest studies of judging that can be traced to the 1800s (Bejar 2012) saw the composite nature of judgments as central to how they are formed. Seminal early work came from Gustav Fechner, writing in 1863 about how an assessment of a work of art is conducted, based on the human ability to discriminate between shades of colour and similar phenomena. The ability is presumably activated by a particular combination of cues and stimuli stemming from the discernible features of the work of art being observed.

This initial research, together with his work that followed, such as Fechner (1897), is often credited with laying the groundwork for a number of different approaches to accounting for aesthetic judgments (e.g., Leder et al. 2004). The crux of this understanding is the proposition that observable, concrete features found in works of art form a basis for accounting for aesthetic evaluations (Bejar 2012). This direction of inquiry proved to be the catalyst for many other studies (Prall 1929; Peters 1942; Carlson 1981; and more), which together constitute norm-based or analytical (Sadler 1989, 129) accounts of the process of making judgments.

---

1 The assessment context generally includes, on the one hand, the rater’s own expectations, the relevant rating scale, and other externally prescribed evaluation-relevant criteria, and, on the other hand, the text being observed, all constituting the comparative framework in which a writing performance is rated.
The common core of analytical models of judging lies in understanding this process as an interaction between the rater as the information processor and sets of distinctive features serving as stimuli. One of the most significant models to emerge in this line of argumentation was the lens model (Brunswik 1952). Later providing the foundation for a major new perspective in psychology known as Probabilistic Functionalism (Postman and Tolman 1959), the lens model of judgments has been adapted to account for many research contexts in the humanities and social sciences (Engelhard, Wang, and Wind 2018).

Over half a century of applying and improving the model has resulted in the development of a general cues-based schema for describing human perception on the basis of the original model. In the centre of the process there is the latent variable that is being evaluated (Everitt 1984). An individual rater utilizes a set of related cues in a more or less unique fashion and provides a central response on the basis of the observation (Hammond 1955). In addition to the lens model itself and its different instantiations used to underpin analytic theories of judgments, further conceptual representations supporting the same idea can also be found. These include decision trees (Braun, Bejar, and Williamson 2006), the normative rater cognition model (Bejar 2011), and also the implicit models outlined by Suto and Greatorex (2006) and Crisp (2012).

Ultimately, the common thread that runs through the analytic accounts of how judgments are reached lies in envisaging the judging process as reducible to a formula comprising a number of formal criteria. This sweeping general claim and the different models that espoused it can be seen as empirically backed up by various studies. One such example of findings supporting this claim is the overlap discovered between rating scales, on the one hand, and the text features and qualities noted by raters as assessment-relevant, on the other (Milanovic, Saville, and Shuhong 1996). Additional examples include the many recorded cases where there is considerable overlap between raters agreeing on the quality of texts without formal recourse to rating scales (Morgan 1996).

When pointing out the downsides of this kind of modelling, researchers often state that the argumentation propping it up is limited to an excessive degree to the over-simplified linear weighing up of attributes (Bejar 2012, 3). Such a reduced process does not correspond to the complexity inherent in the rating task, which is closely bound up to an array of associated skills and abilities and varies subtly under different contexts of assessment. A modelling paradigm that is more composite in its approach thus seems to be called for. This complexity, with the problems it poses for the analytical models, was first addressed in the mid-20th century in studies revolving around holistic scoring (Coward 1952; Diederich, French, and Carlton 1961; Godshalk, Swineford, and Coffman 1966; Myers, McConville, and Coffman 1966; and more).

The apparently counterintuitive idea of a rater assigning a score to a writing performance holistically, based on a general impression, was not easy to accept as a substitute for the image of a deliberate and constructed analysis of a performance (Elliot 2005, 160). However, research showed that a less deliberate analysis reduces the cognitive effort and potentially increases rater agreement (Bejar 2012, 3). With this came the firm realization that much more is indeed involved in the judging process than an analytical model can
capture. This insight gave rise to the development of configurational theoretical accounts of judgments (Kaplan 1964, 187).

Configurational theories of judgments see the act of judging as entailing, to a far greater extent than recognized by any analytical model, a number of intuitive processes, which are, in turn, dependent on a multifaceted set of anything but linearly interplaying factors. This density can be explained by the fact that, in purely psychological terms, judging is a process of bounded rationality (Gigerenzer and Selten 2002; March 1994; Rubinstein 1997; Simon 1957). This means that human evaluation procedures are seriously impacted by our limited information processing capacity. This can be exacerbated by having to arrive at categorization decisions under conditions of uncertainty, such as is the case when there are not enough tangible criteria to guide the process. This is certainly true for assessment of writing and the rating scales used for this purpose.

However, in the context of researching the judging processes at work in the field of assessing writing performances, statements such as “I know a [2:1] when I see one” (Ecclestone 2001, 301) and the existence of significant levels of agreement over the quality of a text, even without any recourse to common rating scales or shared rater training (Morgan 1996; Wolf 1995), seem to support the view that judgments are not solely configurational in nature, either. Therefore, there is evidence that the process of judgment seems to be rule-governed at its core and activated by certain tangible stimuli resident in observed texts and expected by the operating assessment context and the operating raters, while there still being an unformalizable complexity fuelling the holistic dimension of judging.

Such reasoning leads to a consideration that there is a need for an understanding of judgments which conceptualizes the judging process invoked in rating a writing performance as seemingly intangible on the surface, but nevertheless structured in terms of deep processing, as a hybrid model. Such a model of judging would take equally from both the analytic and configurational approaches, and would follow the analytic conceptualization of judgments in as much as it would accept the existence of assessment-relevant stimuli in the observed responses (i.e., writing performances) as the formal basis of judgments. It would follow the configurational conceptualization by allowing for the interaction between raters and the assessment-relevant stimuli – as the formal basis of judgments – to be much more complex than can be expressed by simple counts or, most likely, by any kind of linear representation. The hybrid model would also allow for a complex interaction of other factors and raters, beyond the rating-relevant features alone,² thus acknowledging many potentially construct-irrelevant stimuli that may bring bias into the whole process. The plausibility of such hybridity in the process of judging is also vividly illustrated when judgments are revealed as being largely preconscious in nature.

² Such potentially construct-irrelevant influences can result from, for instance, the problems that rating scales bring in terms of vagueness of descriptors (Knoch 2011). Rater backgrounds may also lead to undesired score variability, together with the stylistic differences in the cognitive approaches of the raters (Scott and Bruce 1995). Other sources of bias can be rater strictness or potentially discriminatory behavior (Park, Chen, and Holtzman 2015). All of these can have been found to affect the decision-making and the rating of writing performances (Thunholm 2004).
2 Judgments as Preconscious Processes

Different studies within the educational field have tried to account experimentally for the cognitive nature of judging (Cooper and Odell 1977; Crisp 2010; Faigley 1989; Grainger, Purnell, and Zipf 2008; Tigelaar et al. 2005; and more). Most notably, there is the dual-processing theory of judgments proposed by Suto and Greatorex (2008). It focuses on registering the intensity of cognitive processing and follows from the awareness of the decision-making process which invariably takes place within the constraints of the limited human capacity to process information. Suto and Greatorex suggest that judgments can be classified as falling into one or other of two categories (2008, 215):

SYSTEM 1 judgments: judgments in this category are quick, associative, and often intuitive. The judging actions undertaken in this manner are preconscious and effortless. They require little to no recourse to rational thought and happen rapidly and in parallel. Examples of such judgments can be found in common statements such as the already mentioned “I know a [2:1] when I see one” (Ecclestone 2001, 301); and

SYSTEM 2 judgments: judgments in this latter category are slow, rule-governed, and raters are aware of employing the relevant rules. Such judgments are systematic, they require a lot of cognitive effort, and are, in principle, fully consciously controlled by the rater. However, it is important to note that for individual raters SYSTEM 2 actions tend to become SYSTEM 1 actions with increased experience. This is evident in the attested fact that the more experienced the raters are, the less likely they are to overtly follow prescribed rules (Wolf 1995).

The existence of two broad categories for the types of mental processes involved in the act of rating (judging), one more intuitive and the other more rational, is also indicated by a number of similar studies, such as Milanovic, Saville and Shuhong (1996). It suggests the presence of an initial skimming phase, when raters assign a provisional internal score, followed by an in-depth phase, in which raters come back to their own initial score and double-check. This latter phase basically includes comparing the original intuitive judgment based on first impressions with prescribed rating criteria and thus serves as an internal monitoring or an internal standardization procedure. An analogous division is also posited by Cooksey, Freebody and Wyatt-Smith (2007), Ecclestone (2001), Keren and Teigen (2004), Lumley (2005), and Shirazi (2012), among others. In short, the focus of this line of research is on clearly differentiating between the rational thought processes and the intuitive responses involved in judging (Brooks 2009).

The challenge facing researchers investigating the judging process is one of accounting for the process of making tacit knowledge explicit (Eraut 2000, 119). Linking judgments to heuristics is therefore also in line with this kind of research. As a process, heuristics reduce cognitive complexity and represent a means of facilitating a mental shortcut (Gilovich and Griffin 2002, 3). Heuristics function by reducing uncertainty via a reference to previously established “anchors” (Mussweiler and Englich 2005; Tversky and Kahneman 1992). External pressures from the volume of work, time constraints, and the like are often additional stress factors in evaluation. Therefore, it is easy to understand the tendency of raters to either deliberately or pre-cognitively create strategies which facilitate shortcuts in order to complete the judgment
process in the most resource-economical way possible (Gigerenzer and Goldstein 1996). The ultimate goal, motivated both from within, cognitively, and from without, as resource-based pressure, seems to be to minimize the intellectual effort and investment of resources but to still get the job done (Krosnick 1991; Quinlan, Higgins, and Wolff 2009; Simon 1957).

What can be concluded from this discussion is that arriving at judgments can be seen as a combination of rapid, preconscious processes (SYSTEM 1) and conscious rule-governed processes (SYSTEM 2). The central tendency, nevertheless, is for the structured explicit processes to become intuitive, an evolution governed by limited cognitive handling capacity and driven by the increasing experience and expertise of the raters. Less experienced raters tend to focus more consciously on specific quality-relevant features. With experience, however, they tend to approach texts in a more pre-cognitive manner and to ultimately adopt a more holistic approach (Furneaux and Rignall 2007).

Nevertheless, in the migration from conscious, more analytic processing (SYSTEM 2) to the more preconscious, holistic approach (configurational, SYSTEM 1), it seems that only the manner of processing is changed (from slow and calculated to fast and intuitive) rather than what actually drives the process. There remains something largely consistent underlying congruent judgments made by different raters. This seems to indicate that the psychological perception of judgments also confirms the existence of both analytic and a configurational side to the judging process, in congruence with the discussion offered above. A similar conceptualization of judging also appears to stem from the perception of judgments as complex comparison procedures.

3  Judgments as Comparisons

That the forming of judgments, with the intricate processing of information which this entails, is suitable to be imagined as involving a complex procedure of comparison (Laming 2004, 9) was evident even in the earliest models of judging, such as the previously referenced lens model (Brunswik 1955). Here, comparisons were being made between certain activated mental representations and certain salient features of the observed phenomenon. In the context of rating writing performances, judgments can, in this respect and as a starting point, be conceptualized as essentially reading comprehension procedures. The meaning of the text is actively constructed, based on the interaction between the readers’ knowledge and the content supplied by the text (Johnson-Laird 1983, 402).

Cumming (1990, 33) similarly suggests the existence of two general clusters of rating attention when it comes to writing assessment, namely one concerned with reading and another, interpretive one. The actual decision-making occurs as a mode of retrieving relevant information from memory and comparing it to the meaning constructed on the basis of the observable textual input (Baker 2012, 226). In this respect, rating can be conceptualized as the process of comparing the knowledge structures built on the basis of the observed performance with the rater’s own internally activated corresponding mental representations of the self-same performance (Crisp 2008), generated by the rater’s own expectations and largely motivated by the expectations of the assessment context.
Even when the description of the rating process is not confined to accounting for only a few general steps, but instead is more finely grained, it can still be conceptualized in terms of making comparisons. For example, Bejar describes the scoring procedure as a sequence of several phases (2012, 5):

1. the rater first reads the writing performance and produces a mental representation of it, a process which takes the form of a comparison between the context-conditioned hypothetical “ideal” instantiation of the observed text and the observed text itself;
2. the rater then weights the previously formed mental representation in relation to a mental scoring rubric. The mental scoring rubric represents the rater expectations in more explicit form, including potentially internalized rating criteria; and
3. the rater assigns the performance to a score category, possibly after a self-monitoring phase which can include a more conscious reference to the externally prescribed criteria, referenced by experienced raters most commonly for the purposes of uniformity of expression rather than for modifying the impression already formed.

Myford and Wolfe view judging as a process of information acquisition and processing, punctuated by a number of rater functions (2003, 37). These include: (1) the rater functioning as an information processor; (2) the rater retrieving information from memory storage; and, ultimately, (3) the rater combining, weighting, and integrating the information to draw rating inferences. Wiliam also lists several comparands, including criteria, constructs, self, and norms (1992, 17). DeRemer conceptualizes the rating process as a problem-solving exercise (1998, 27), including (1) the simple recognition task elaboration (i.e., general impression scoring); (2) the rubric searching elaboration (i.e., rubric-based evaluation); and (3) the complex recognition elaboration (i.e., text-based evaluation). Finally, synthesizing a myriad of accounts, Gauthier, St-Onge and Tavares list seven general mechanisms involved in making judgments, which can be grouped into the three phases of (1) observation (reading); (2) processing; and (3) integration (2016, 511). These mechanisms comprise:

1. generating automatic impressions;
2. formulating high-level inferences;
3. focusing on different dimensions of the competencies of interest;
4. categorizing through well-developed schemata, based on personal concepts of the given competence, comparison with various examples, and also task and context specificity;
5. weighting and synthesizing information in a number of different ways;
6. production of mature judgements; and
7. translating narrative judgments into scales.

On the part of the rater, this comparison process can be affected by several additional factors. These include cognitive constraints (Simon 1957, 227), prior knowledge and experience (Newell, Lagnado, and Shanks 2007), and possibly a number of potentially construct-irrelevant influences (i.e., rater effects, see Eckes (2023)). The cognitive constraints, as already noted, are the limiting factors in human cognition that are responsible for ultimately making
judgments, to a large extent, the result of preconscious intuitive processes. Prior knowledge and experience are embodied, as a combination with externally prescribed criteria, into the fuzzy hypothetical “ideal” form a particular performance should take in a particular context following a particular prompt.

This firstly corresponds to the mental representation that is activated rater-internally, existing in potentia in the rater’s mind and instantiated only as an authoritative reconstruction of the observed text, embodying the rater’s expectations. This rater-internal point of comparison can be further influenced by various external and co-opted rating criteria that a rater has been asked to use, and through experience, has potentially internalized and learned to apply (Suto and Greatorex 2008, 215). The fact is that, in the course of the migration from self-aware to preconscious judging, the externally prescribed rating criteria can undergo change and become adapted to fit the already existing rater expectations (Brooks 2009, 12). Finally, rater expectations, commonly incorporating external rating criteria, can be further affected by various factors such as cultural background, rater strictness, and different sources of bias (Hay and Macdonald 2008, 159).

That making comparisons is central to the nature of judgments is also revealed by studies looking into the weighting patterns that raters exhibit. Cumming (1990, 47) shows that among 28 identified types of rater behaviour, raters, overall, focus on comparisons based on self-monitoring on the one hand and rhetorical, ideational, and language content on the other. Chalhoub-Deville, in her discussion of deriving assessment scales for oral performances, similarly identified grammar-pronunciation, creativity in presenting information, and amount of detail provided as relevant comparands (1995, 48). Cai (2015) found raters putting emphasis on either form or content, or on achieving a balance between the two. Eckes (2008) categorizes raters according to the weight they attach to syntax, correctness, fluency, non-fluency, and non-argumentation. Baker (2012), in turn, classifies raters as regards their emphasis style being rational, intuitive, dependent, avoidant, and spontaneous.

In summary, it can be concluded that conceptualizing judgments as comparisons also seems to confirm the existence of concrete features in texts which serve as stimuli for rater cognition and, ultimately, for categorization decisions. Central to the process of comparison are, in fact, discrete aspect of texts which are recognized by raters as being in line or not in line with their (assessment) expectations relevant for a particular educational/assessment context and relative to a particular performance being observed. It is not surprising, therefore, that much research effort has been invested into describing what these discrete aspects of texts may comprise.

4 Assessment-Relevant Stimuli Responsible for Triggering Micro-Judgments – Errors as Prototypical Cues

Common to the presented conceptualizations of the dual nature of judgements (i.e., configurational vs. analytical accounts, the dual-processing view on the underlying cognition, and the conception of judgments as comparisons) is their direct or indirect focus on trying to pinpoint concrete types of assessment-relevant text cues and stimuli by which the raters may be influenced when assessing writing. This has, in fact, always been one of the central questions in rater cognition research.
For example, Milanovic, Saville and Shuhong identified 11 essay elements constituting rater focus, with different weighting attached to each by the informants, whereby these included *grammar, communicative effectiveness, content, length, legibility, spelling, structure, task realization, tone, punctuation,* and *vocabulary* (1996, 106). Similarly, the study conducted by Sakyi identified four concrete focus points, including *errors* in the text, *essay topic* and *presentation of ideas*, raters’ personal reactions to the text, and *prescribed scoring criteria* (2000, 140). Cumming, Kantor and Powers report clustering of rater focus regarding three broad categories of text features, including *rhetorical organization, expression of ideas,* and *accuracy and fluency* (2001, 3).

The general conclusion from these and other similar studies, stemming mainly from interviews and think aloud protocols, seems to be that raters essentially focus on three major qualities of a text, which can be loosely termed correctness, content representation, and relation to the task. This has indeed been recognized by the rich field of CAF research (Skehan 2009), with studies stipulating *complexity* (as expectations of mature language performance, see DeKeyser (2008)), *accuracy* (as expectations of norm-adequacy, as in Xie (2019)), *fluency* (as expectations of smoothness or ease of speech or writing, see Hilton (2008)), and *functional adequacy* (as expectations of task achievement, in Kuiken and Vedder (2017)), as the main cornerstones of language competence.

What such studies generally do not focus on, however, is that the crucial aspect of assigning assessment-relevance to any text feature involves making the distinction of whether it has been performed successfully and in alignment with the operating assessment expectations or not (Dobrić 2024). It is not enough just to catalogue the different types of features that can be assessment-relevant. The manner of identifying the direction and magnitude of their assessment-relevance when observed in action, appearing in a concrete performance, is also necessary to consider. For instance, every time a text feature is recognized by a rater as not performed as expected, regardless of whether this is in relation to accuracy, complexity, fluency, or functional adequacy (thus being highlighted as *inaccurate* and/or *inappropriate*), a certain degree of *assessment-negative* value is usually recognized for it, too. In other words, we have a case of an error being registered (i.e., micro-judged), triggering on the instance of an assessment-relevant stimulus (usually in a negative direction) on the side of the rater.

Errors have been referenced for generations as text cues relevant to writing assessment, typically found underlined using the dreaded red pen and often implied as constituting much of the basis of the grade received on a writing task. An error, in this sense, is understood as any text feature which does not conform to rater expectations relevant for one assessment context. This includes both the more traditional understanding of errors as related to correctness and *non-norm adequacy* (Richards 1974) and, far beyond any formalizable norm, those related to *inappropriateness*, relative to performance expected in fulfilment of a specific assessment task in a specific setting.

Despite this, “error counting” approaches to the rating of writing performance are generally viewed very negatively by the applied linguistics field, and it is not difficult to see why. Namely, an emphasis solely on errors in the assessment of productive language skills is likely to be counter-productive, misrepresentative, and demotivational. A “good” text entails significantly
more than it only being “error-free”. Still, reference to errors as important assessment-relevant cues and stimuli persists in educational practice, for several reasons (Dobrić and Sigott 2023).

Firstly, an attention to errors provides raters with an intuitively transparent set of assessment criteria, generally offering much more structure and clarity than the often rather vague rating scale descriptors. Secondly, the usefulness of errors also stems from them being conceptualized as indicators of problems in the learning process, (Dobrić 2015; Dulay, Burt, and Krashen 1982; Havranek 2002; James 1998; Lengo 1995; Lennon 1991; Swan and Smith 2001). This is what makes errors so indispensable for the provision of constructive feedback (Sigott et al. 2019).

For all these reasons, it is relatively safe to claim that errors represent a prototypical case of a type of assessment-relevant set of text features generally influential when making judgments of writing performance (Dobrić and Sigott 2014). They are one of the key features that facilitate the comparison between the rater's mental representation of an observed text and the observed text itself. Therefore, the manner in which errors, for whose recognition a significant amount of comparison (i.e., judging) is required in the first place, go on to help facilitate this process of complex comparison is key to understanding the nature of their contribution to forming judgments (and, for that matter, of the contribution of any type of conceptually comparable assessment-relevant text cues or stimuli one may propose beyond errors).

5 Micro-Judgments into Macro-Judgments

The manner in which errors, as a well-established and widely recognized type of assessment-relevant text features, can be understood as contributing to judgments has been illustrated in the manner of how judgments largely function as fast, preconscious comparisons. Errors, as one participant in this complex comparison procedure, tend to indicate “undesirable” divergences between, on the one hand, the written performance that is actually observed and, on the other hand, the written performance expected by the rater and the surrounding assessment context. This weighting of similarities and differences between rater-internal expectations and rater-external stimuli, or in other words, the comparison between the hypothetical ideal version of a text (for a set context) and the self-same text as actually performed, results in a “vector of similarities” being constructed using these two broad comparands (Bejar 2012, 5).

This vector is a result of a reconciliation taking place between the rater’s overall impression of a text (as reflected in their expectations), the specific yet often hard to pinpoint, assessment-relevant text features (often even explicitly recorded for feedback purposes), and, finally, any given (external or internalized) rating scale criteria (Zhang 2016). Ultimately, the vector of similarities ends up representing the overall judgment of a text and directs the predisposition,

---

4 In truth, the appeal of including errors in the assessment process may go even further, perhaps answering to the human instinct to associate any kind of evaluation first and foremost with pointing out faults. This is observable in many human-based evaluation practices, such as academic project reviews or professional performance evaluations.

5 Explicit attention should once more at this point be drawn both to the fact that this process of comparison can be additionally influenced by various sources of potential bias, and that this process is anything but as linear or straightforward as may be inadvertently suggested by the narrative.
in the mind of the rater, to classify a performance as belonging to a particular quality-related category by assigning it a rating.

The novel insight important to highlight at this point – stemming from how errors have been accounted for as one set of text features raters themselves often report as relevant when assessing, and from how judging has been described as largely a preconscious procedure – is that the composition of the vector of similarity depends, by default and to a large extent, on individual instances of comparison on the text feature level, building up its geometry as coordinates. Namely, as Figure 1 shows, individual instances of comparison between text features (as assessment-relevant stimuli) expected and text features observed cumulatively (and, again, non-linearly) build up into an overall comparison being formed between a writing performance expected and writing performance observed. They thus construct the said vector of similarity which is ultimately translated into a rating.

The effects of any assessment-relevant text cue or stimulus are relative to the specific assessment expectations operating in a particular context (including here also the relevant language norm, see Gilquin (2022)). This makes them essentially latent phenomena, with each individual instance of comparison on a text-feature level, key to recognizing any error, requiring a judgment to take place on the side of the rater. A micro-judgment, as termed and understood here, thus involves any one individual text feature in a writing performance being (mostly preconsciously) recognized as assessment-relevant (alongside its negative or positive direction and magnitude of effect) by a human rater. It is flagged thus for the rater’s own benefit and for the purpose of quality categorization by virtue of the direction and extent of its accuracy and/or acceptability (or lack thereof) in a specific context.

**Figure 1.** Simplified illustration of the role of errors as triggers for micro-judgments.

What makes errors ideal exemplification candidates in this case is that, as indicated earlier, the process of recognizing one text feature as an error (similar, in fact, to recognizing any text feature as assessment-relevant) is the same in structure as described previously in relation to forming an overall judgment of a writing performance, as illustrated in Figure 1. The process
requires the rater to judge the conformity of any given text features in terms of its correctness and acceptability, and it entails all of the complexity discussed in the previous sections, only on a smaller scale, involving one assessment-relevant text feature at a time.

Correctness, as one coordinate of rater expectations relevant for identifying errors, references the more codifiable and, hence, the more firmly codified aspects of language performance (i.e., language norm and standards). It is usually incorporated into the assessment expectations of all educational contexts involving language. Inappropriateness, on the other hand, does not relate to any firmly codified (or even codifiable) aspects of language performance, but rather to often gradable judgments made as to the degree to which particular text features conform to particular operating (general and specific, text-internal and text-external) assessment expectations (Dobrić et al. 2021).

For instance, if we take the “After buying the book, he sat to read her” example sentence used in Figure 1 to illustrate the ‘TEXT FEATURE OBSERVED’ step of a micro-judgment being triggered, the manner in which the text feature her is detected in it as an assessment-relevant cue of negative direction (i.e., as an error) has to do with it not fitting to what a rater would expect when guided by expectation of English language competence relevant to anything but lowest levels of proficiency. The expectation, pointed out by the “After buying the book, he sat to read it” is one of suitable anaphoric reference made by the underlined pronoun and is motivated by both appropriateness and correctness as reference points. The micro-judgment that takes place is mirrored in the rater engaging in the comparison between the expected and perceived and deciding on how to interpret any discrepancy. In this case, this would take the direction of her being judged to be an error.

This and the other examples provided in Figure 1 solidify the claim that an overall judgment of a writing performance, which has been, in the light of this argumentation, termed a macro-judgment, is comprised of accumulated and in some way combined individual micro-judgments, when it comes to errors as stimuli at the very least. This accumulation process is, as indicated above, complex and involves an as of yet unmapped scheme of non-linear “counting and combining” of the micro-judgments into a macro-judgment. Nevertheless, the research indicates that the initial result of this process of macro-judging is also achieved preconsciously and generally takes the form of a fairly rapid decision on the categorization of the text (taking place as a SYSTEM 1 judgment, as in Suto and Greatorex (2008)). This is despite the fact that the final step in the process is one potentially involving some conscious deliberating, as a self-monitoring stage, during which external criteria (i.e., the rating scale) may be additionally consulted, in a more conscious and self-aware manner, prior to awarding the rating. In addition, as also already underlined several times, different biasing and presumably not construct-relevant factors may also influence the construction and direction of the relevant vector of similarity.

Hence, as Figure 1 illustrates in a necessarily simplified manner, in order to recognize a writing performance being assessed as conforming or not conforming (and to what degree) to the expectations operating in one assessment context, raters first judge each and every text feature in it (individually and in combination) to the same extent of potential conformity. This, as many studies have indicated, usually happens preconsciously (especially with experienced
raters), leaving the manner in which these lower-level micro-judgments are combined into an overall macro-judgment still largely opaque. When it comes to errors as a prominent and traditionally influential type of assessment-relevant features, the micro-judgments occur at the moment in which any text feature expected to appear in the text observed is not found performed in the anticipated (i.e., correct and/or appropriate) fashion or is perhaps missing.

Other assessment-relevant factors identified in the different studies listed before, besides errors, presuppose additional types of assessment-relevant text features being postulated, presumably constituting a group that can be interpreted as assessment-positive, i.e., as diametrically opposite to errors which are assessment-negative in effects (Dobrić 2023). These can be expected – themselves by default also representing latent phenomena – to also assert influence via the micro-judgments they trigger, only in the direction contrary to that of errors (i.e., with a positive value, steering raters towards awarding a higher rather than a lower rating).

6 Errors and Judgments – A Case Study

Empirical backing for the assertion of the prototypical role of errors in forming micro-judgments which consequently inform (together with other kinds of assessment-relevant text features) the macro-judgments of writing performances offered above can be found in a recently published study by Dobrić (2024). The research, focusing on the effects of errors on rating writing performance, demonstrates how errors explained roughly 50% of overall score variability in a high-stakes EFL context. This was found despite the fact that the rating criteria (i.e., the rating scale) employed in the given context actually directed the raters towards a focus on assessment-positive aspects of texts (rather than the assessment-negative qualities of texts that are represented as errors). Table 1 briefly recaps the reported results.

Table 1. Pseudo R² coefficients (Nagelkerke 1991)⁶ reported in the model, representing the overall effects of errors on the different rating dimensions of the employed analytic rating scale (Dobrić 2024).

<table>
<thead>
<tr>
<th>Ratings</th>
<th>Nagelkerke Pseudo R-Square</th>
<th>Variability of Ratings Accounted For</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>0.358</td>
<td>36%</td>
</tr>
<tr>
<td>Grammar</td>
<td>0.617</td>
<td>62%</td>
</tr>
<tr>
<td>Textual Competence</td>
<td>0.569</td>
<td>57%</td>
</tr>
<tr>
<td>Aggregated ratings</td>
<td>0.512</td>
<td>51%</td>
</tr>
<tr>
<td>average</td>
<td>0.514</td>
<td>51.5%</td>
</tr>
</tbody>
</table>

What the study demonstrates, in terms of how judgments were formed in the case of the high-stakes writing exam observed in the study, is that a concrete set of distinctive text features (which can be argued as assessment-relevant) could be empirically brought into correlation with the independently formed rater judgments (expressed as grades). In other words, even

⁶ Pseudo R² coefficient is a supplementary measurement used to demonstrate the compound effect of independent (i.e., predictor) variables in explaining the variance of the dependent (i.e., predicted) variable on the whole, emulating the more familiar R² coefficient from linear regression (Veall and Zimmermann 1996).
though the raters who originally processed the writing performances analysed in Dobrić (2024) received both overt and covert instruction (i.e., by means of rater training and via the rating scale employed for that purpose) in how to engage in judgments (i.e., employ a more configurational rather than analytical approach), their decisions could still be quantitatively (i.e., by means of logistic regression) linked, in a causal manner, to individual features of the rated texts (i.e., in this case to errors).

As already argued above, each of the individual features of a text comprising an instance of an error demanded a judgment to be made as to it actually constituting an error in the first place. Considering that the analysed writing exam took place at a university department of English and involved C1–C2 expectations of L2 English proficiency, there were many errors which demanded decisions being made in terms of their appropriateness rather than of their correctness (Dobrić and Sigott 2023). What this lucidly evokes is the image of micro-judgments, triggered by errors, cumulatively informing macro-judgments, expressed as grades, as illustrated in the section above. In the case of Dobrić (2024), this was attested to be at the rate of between 36% and 62% of the variability of the macro-judgments (as indicated by the pseudo R² values provided in Table 1) likely informed by error-based micro-judgments.

What is more, the fact that it was around 50% of the variability of the grades observed in the examined high-stakes university writing exam covered by Dobrić (2024) indicates that something else is responsible for the unexplained portion of the variability of the self-same grades. It is assumed but not yet empirically demonstrated that the proposed assessment-positive text features (Dobrić 2023) may play a significant role in this respect, possibly responsible for macro-judgments in a similar way as errors when triggering micro-judgments, only of a different (i.e., opposite) assessment value and direction.

7 Consequences of the Findings for Understanding Writing Assessment

Even though the exact manner in which micro-judgments (and possibly other factors) interact and come together to form overall macro-judgments (ultimately filtered through rating scales into ratings) is still largely unaccounted for (and very difficult to formalize, being of both analytical and configurational nature), the realization that distinctive cues and stimuli found in assessed writing performances also require the activation of a judgment process is a noteworthy one. It is not only relevant in terms of contributing to our understanding of rater cognition but is also important, in a practical sense, in terms of rater training and the construction of rating scales.

Starting with rater cognition, the realization that an overall judgment of a writing performance is comprised of many, mostly preconsciously made, smaller-scale judgments is significant because it reveals a whole new level at which subjectivity may be affecting measurements of writing competence, thus highlighting the need for accounting for this in some manner (similarly to how rater effects are accounted for at the level of macro-judgments and scoring, see Eckes (2023)). Namely, just the example of errors as one possible type of assessment-relevant text features likely responsible for triggering micro-judgments indicates the issue of gradience.
intrinsically to recognizing the majority of error types and of their potential assessment-relevance and, by extension, the amount of potential subjectivity involved (as often demonstrated in by studies looking into inter-annotator agreement, see Sigott, Cesnik, and Dobrić (2016)).

The issues of gradience and subjectivity are only aggravated when other types of assessment-relevant text features that may be postulated are considered, such as the hypothetical assessment-positive ones (Dobrić 2023). Highlighting concretely what makes one writing performance “good” (i.e., acceptable or even extraordinarily acceptable) in terms of concrete text features and qualities related to one assessment context is, in contrast to errors – which do at some level also respond to the much more formalized correctness as a reference point – almost exclusively an act resident in judging. This is because determining assessment-positiveness involves solely deciding on the levels of adequacy (or excellence) of individual aspects of a text for the set context and not of correctness at all, possibly adding even more bias and potential for measurement error to the process.

This is why it is crucial to acknowledge the existence of an additional underlying level of potential subjectivity involved in rating writing (Dobrić 2018) and cast it into methodological solutions which would allow us to account for it when it comes to the rating, rater effects, and potential measurement errors it can produce. Similarly, it is also important to cast the outlined findings into rating practice by making the raters and rating criteria more sensitive to them (Cushing Weigle 1994; Dobrić et al. 2021). In other words, while methods for detecting and measuring rater effects in performance assessments have for a long time been a fixture of psychometric research on rating quality in the assessment of writing (Eckes 2015), the diverse methodology stops at the level of the text as the measured phenomenon and the macro-judgment as the relevant process. The discussion so far has, however, shown that there is a need to account for measurement errors being produced below the text as a whole, in both testing research and practice. One step towards doing so is to utilize methods designed for describing annotator behaviour in learner corpus coding (Dobrić 2022). The myriad of methods devised for measuring intra- and inter-annotator agreement, usually revolving in some way around Kappa coefficients (McHugh 2012), could prove very useful for revealing the rater effects potentially present at text feature levels. This may include, similarly to those found at the text level, the raters when micro-judging individual textual features showing evidence of inaccuracy, illusory halo, severity/leniency, central tendency/extremity, and so on.

Then, in rater training, more emphasis needs to be placed on having raters become more aware of concrete text features and textual qualities that may be influencing their impression-making when it comes to writing performance (Cushing Weigle 1994). This can take the form of different exercises focused on identifying assessment-relevant cues and stimuli in example texts and then accounting for why they may be signifying relevance in the first place. Furthermore, emphasis should be placed on the different reference points necessary for the raters to employ in order to recognize a text feature or quality as assessment-relevant, such as the norm, the rating criteria, the authoritative reconstruction, and similar.

---

7 Revolving largely around many-facet Rasch measurements, they are designed to detect and somehow account for measurement error human raters bring to the table when assessing writing. These include inaccuracy, illusory halo, severity and/or leniency, central tendency and/or extremity, and more (Wolfe and Song 2016).
Open for further research, and much more controversial than the proposition of the existence of micro-judgments, is the proposition that the manner in which they interact to comprise the macro-judgments (alongside bias as potential additional factors) can be analytically presented. This is precisely where the most prominent opportunities for study are to be found, especially in terms of natural language processing (NLP), artificial intelligence (AI), and algorithmic representations of rater cognition. Dismantling the complex cognitive process in which the vector of similarity between the expected writing performance and the observed writing performance is actually created in the rater’s mind and doing so in a way which would not be solely reliant on raters’ own testimonies (fraught with problems of reactivity and veridicality of informant responses, see van Someren, Barnard, and Sandberg (1994)), would be the next natural and necessary step to take in our understanding of what raters do when they assess writing and in our attempt to formalize it.

In reference to these open questions, additional studies would also be welcome in the direction of investigating further the different types of assessment-relevant text features other than errors that can be demonstrated as also responsible for triggering macro-judgments. Moreover, studies into methods which would allow for a better understanding of rater effects and of measurement errors potentially occurring at the text feature and micro-judgment levels, would also prove a promising research direction to take. Finally, studies which focused on cataloguing assessment-relevant features commonly appearing in specific contexts and using the findings for improving rating scales and rater criteria in the self-same settings, would also be of major academic and practical significance.

References


