Studia Anglica Posnaniensia 52(2), 2017
doi: 10.1515/stap-2017-0006

DIFFERENCES ACROSS LEVELS IN THE LANGUAGE OF AGENCY AND ABILITY IN RATING SCALES FOR LARGE-SCALE SECOND LANGUAGE WRITING ASSESSMENTS

SALENA SAMPSON ANDERSON*

Valparaiso University

ABSTRACT

While large-scale language and writing assessments benefit from a wealth of literature on the reliability and validity of specific tests and rating procedures, there is comparatively less literature that explores the specific language of second language writing rubrics. This paper provides an analysis of the language of performance descriptors for the public versions of the TOEFL and IELTS writing assessment rubrics, with a focus on linguistic agency encoded by agentive verbs and language of ability encoded by modal verbs *can* and *cannot.* While the IELTS rubrics feature more agentive verbs than the TOEFL rubrics, both pairs of rubrics feature uneven syntax across the band or score descriptors with either more agentive verbs for the highest scores, more nominalization for the lowest scores, or language of ability exclusively in the lowest scores. These patterns mirror similar patterns in the language of college-level classroom-based writing rubrics, but they differ from patterns seen in performance descriptors for some large-scale admissions tests. It is argued that the lack of syntactic congruity across performance descriptors in the IELTS and TOEFL rubrics may reflect a bias in how actual student performances at different levels are characterized.

Keywords: rating scales; second language writing; writing assessment; performance descriptors; linguistic agency

1. Introduction

While the literature on language testing and writing assessment is rich with studies evaluating the validity and reliability of given assessments, a relatively

* Department of English, 223 Arts and Sciences Building, 1400 Chapel Drive, Valparaiso University, Valparaiso, IN 46383; e-mail: Salena.Anderson@valpo.edu.

smaller body of literature explores the actual language of writing rubrics themselves. In this study, writing rubrics may be understood as demonstrating the same range of features and functions described by Covill (2012): "a list of criteria that are relevant to producing effective writing", generally featuring multiple levels with descriptors, used for rating, placement, instruction, or a combination of these functions.[1] In the case of large-scale assessments of second language writing, such as those that are part of the Test of English as a Foreign Language (TOEFL) or International English Language Testing System (IELTS) exams, raters may experience "tension" between the language of the performance descriptors in the rating scale and their own "intuitive impression" of a given learner text, a tension that is addressed – though not fully resolved – by rater training (Lumley 2002: 246). Lumley goes on to argue that "[r]ather than offering descriptions of the texts, the role of the scale wordings seems to be more one of providing justifications on which the raters can hang their scoring decisions" (Lumley 2002: 266). Further, trained raters are not the only audience for these scales. In an effort to educate test takers, teachers, and schools about their tests, Educational Testing Service (ETS) and Cambridge English Language Assessment, as well as other testing agencies, provide public versions of their rubrics online. Thus the language of these rubrics is available to a general audience, who may use it to prepare for the exam, to borrow or adapt for their own assessments, or to consider in admissions decisions.

Some of this literature that looks at the language of performance descriptors focuses on the content of the rubrics. For instance, Matsuda & Jeffery (2012) analyze (lack of) attention to voice in performance descriptors for writing assessments on national English proficiency tests and standardized tests of college readiness. Jeffery (2009) also provides a content analysis of large-scale writing assessment rubrics and a syntactic analysis of prompts.

The language of performance descriptors, including syntactic structure, is frequently referenced in language and writing assessment literature, but remains a challenging area. The limitations associated with distinguishing between levels exclusively with adjectives and adverbs of degree are frequently mentioned, e.g., Hawkey & Barker (2004: 127), Knoch (2011: 82). Other scholarship on rating scales considers the level of specificity and detail of the descriptors, e.g., Knoch (2007: 121), Brindley (1998), and Upshur & Turner (1995). Specific attention to verbs also appears in the "can do" statements of the *Common European Framework of Reference for Languages* (CEFR), and the language of performance descriptors is again referenced as North (2007) reflects that all the descriptors for the CEFR are "worded in positive terms, even for

[1] See Covill (2012) for a review of scholarly sources discussing key features of writing rubrics that inform her definition.

lower levels" (North 2007: 657). However, the language of performance descriptors in rating scales continues to be an issue for language testing and writing assessment; Kuiken & Vedder (2014: 283) identify "the potentially multiple interpretation and vagueness of some of the scale descriptors" as an area for future research. Alderson et al. (2004) discusses a similar issue in describing levels of reading and listening in language testing.

However, scholarly accounts of the development of rating scales for writing, including the creation and revision of performance descriptors, are limited. In introducing his corpus study of the language of first-year college writing rubrics, Dryer (2013: 5) asserts that "[l]ittle is known about how scales themselves are composed, and few field-tested recommendations for scaling performance categories exist as of this writing". He points out a similar assertion in *Educational Measurement* (Hambleton & Pitoniak 2006: 453), and this observation is echoed more recently by Banerjee et al. (2015).

Dryer (2013) analyzes the grammar and word choice of performance descriptors in university-level writing rubrics, identifying differences between performance descriptors for different levels, e.g., theses that are "demonstrated" at higher levels versus those that are "made" at lower levels (2013: 17). With regard to agency, Dryer argues that "Agentive students disappear in lower performance categories" (2013: 23). Dryer also notes that the mention of readers is "disproportionately present when giving favorable assessments of style (e.g., 'The reader was impressed by this paper's lively tone') and in their strong reactions to the presence of error" (2013: 24). The rubrics analyzed in this study, however, largely "assume native speaker status" (2013: 8).

Focusing on second language (L2) writing rubrics, the present study builds upon and extends the analyses in Dryer (2013), with special attention to linguistic agency, nominalizations, and the language of ability. To do so, this study considers the following questions: (1) To what extent and how are depictions of writer agency and ability encoded linguistically in performance descriptors in the writing rubrics for the TOEFL and IELTS exams? (2) To what extent are there similarities or differences between depictions (or lack thereof) of writer agency and ability in the TOEFL and IELTS writing rubrics as compared with large-scale admissions test writing rubrics (such as the SAT writing rubric) and classroom-based writing rubrics, as described by Dryer (2013)?

Dryer (2013) applies corpus methods to the analysis of the grammar and word choice of performance descriptors in university-level writing rubrics, and the present study addresses a gap in the literature by analyzing the use of a similar set of linguistic features in a different context: large-scale L2 writing assessments. Further, this study employs discourse analytic methods to provide a more detailed picture of the linguistic features associated with textual or writer agency and ability. While corpus methods provide excellent quantitative

data, allowing for possible generalization over a whole corpus, discourse analytic methods are ideal for providing closer analysis of individual texts and how linguistic form and content interact within a specific context.

According to the ETS web site for the TOEFL exam and the IELTS web site for institutions, these exams are recognized by more than 9,000 organizations, with a potentially wide audience for their publicly available rubrics. Crusan (2010: 44) in her discussion of L2 writing assessment in classroom contexts argues the following:

> Clearly, as in the case of high-stakes testing, writing can be reduced to a formula that shapes instruction and perception of writing. But shunning the rubrics used by high-stakes raters will not help. Rather some rubrics are relevant to help students prepare to move past gatekeepers. They can be 'starting points from which we make our own rubrics' (Calkins 1994, cited in Spandel (2006: 19)).

Given the wide recognition of these large-scale, high-stakes assessments, the public nature of their preparatory materials, and the potential for their rubrics to be viewed as 'starting points' for other classroom rubrics, this paper focuses on the language of the public versions of the writing assessment rubrics for these tests.

This paper illustrates how the TOEFL and IELTS scales for writing feature more agentive verbs in performance descriptors for higher scores, with more nominalizations and/or language of ability in descriptors for lower scores. This pattern of agentive language in large-scale second language writing rubrics resembles that discussed by Dryer (2013) for college writing rubrics designed for a more general audience. The IELTS rubrics, however, are distinct from those discussed in Dryer (2013) in that the lowest band descriptors of the IELTS writing rubrics also employ the language of (in)ability. I argue that this pattern of linguistic features has the effect of removing the writer from the description of the writing at certain band levels, in favor of language that may be perceived as more objective. This syntactic variation at the very lowest and highest band and score descriptors also parallels biases found by Schaefer (2008) in the rating of very high or very low scoring essays, and may be unnecessarily disempowering for students with lower scoring essays.

2. Linguistic agency

In a non-theoretical sense, the word *agent* is easily understood as something like the following: "A person who or thing which acts upon someone or something; one who or that which exerts power; the doer of an action." (OED, s.v. *agent*, n.1.a). However, there is considerable work in the field of semantics on the characteristics of agents. One of the most cited theoretical works on thematic roles (i.e., the function of an argument in a specific sentence) and agency,

Dowty (1991: 572) identifies the following as properties of the "agent proto-role", a broad range of verb arguments that function like thematic agents:

> "a. volitional involvement in the event or state
> b. sent[i]ence (and/or perception)
> c. causing an event or change of state in another participant
> d. movement (relative to the position of another participant)
> (e. exists independently of the event named by the verb)"

He argues that "arguments may have different 'degrees of membership' in a role type" (1991: 571); in other words, some arguments of verbs are more agentive than others, depending on how many and what kind of properties from those listed above they exhibit. The sentient doer of a volitional action that has an impact on another thing or person represents a classic and easily recognizable type of agent, e.g., *Sandy* in the sentence "Sandy angrily kicked Pat".

Compared to this theoretical work that seeks to define the qualities of prototypical agents, discourse analytic work attends to linguistic strategies and social reasons for reducing or eliminating agency, such as in rape trial discourse (e.g., Ehrlich 2001). Ehrlich (2001: 36–61) points out how nominalizations, passive voice, and unaccusative constructions can have precisely this effect in discourse, by "mitigating", "diffusing," "obscuring", and "eliminating agency". For instance, she compares sentences like the following: "In the U.S. a man rapes a woman every 6 minutes" and "In the U.S. a woman's rape occurs every 6 minutes", adapted from Henley et al. (1995). This research is part of a larger body of discourse analytic work that treats agency, nominalizations, and related linguistic features. Billig (2008) provides a detailed discussion of research on nominalization and the deletion of agents in influential Critical Discourse Analysis literature, including the seminal work of Fowler et al. (1979).[2]

Importantly, there is also experimental evidence that syntactic choices correlating with more or less linguistic agency encoded in a sentence cause readers to ascribe more or less actual agency to the people and/or entities in a sentence. LaFrance & Hahn (1994) and Henley et al. (1995) have presented this argument, and more recently Fausey & Boroditsky (2010), Fausey et al. (2010), and Chakroff et al. (2015) have presented results from experimental studies showing that more direct and more agentive language causes readers or listeners to attribute more actual agency or blame to an individual who appears as a grammatical agent in a sentence. For instance, in Fausey & Boroditsky (2010),

---

[2] For a more general introduction to questions and methods for discourse analysis with a focus on Critical Discourse Analysis, see Wodak & Meyer (2009), which also provides a discussion of some of the shared dimensions between Critical Discourse Analysis, other discourse studies, and related fields (2009: 2).

after reading one of two versions of a short narrative describing an accidental restaurant fire, subjects were asked about how much blame they attributed to the woman whose napkin started the fire. Subjects who read a version of the story with less agentive language, e.g., "The napkin had ignited," attributed less blame to the woman than subjects who had read a version of the story with more agentive language, e.g., "She had ignited the napkin" (2010: 645). In other words, syntactic choices related to linguistic agency have been shown to inform readers' perceptions of individuals' actual agency.

In the context of language assessment, agency and depictions of agency may inform raters' judgments of linguistic competence. In researching oral language proficiency assessment, Morales & Lee (2015) further explore the social complexities of agency. Referencing the discussions of agency in Bucholtz & Hall (2005) and Ahearn (2001), Morales & Lee caution against understandings of agency that "equate it with autonomy, intentionality, or free will" (2015: 34). They also appeal to "Duranti's (2004) notion of agency, particularly his argument that agents' actions have consequences for themselves or others and are the object of various kinds of evaluation, including the evaluation of one's linguistic competence" (2015: 34). As varying levels of linguistic agency in a text influence a reader's perceptions of the actual agency of the doer (cf. Fausey & Boroditsky 2010, Fausey et al. 2010, and Chakroff et al. 2015), the language of these rubrics, which varies in terms of grammatical agency across descriptors for different levels, may inform raters', teachers', administrators', or students' perceptions of writers as more or less agentive. The language of these rubrics, featuring both agentive language and the language of ability, attempts to comment directly on writer agency and ability, rather than actual performance.

3. Methods

3.1. Discourse analytic approach

Employing discourse analytic methods, this study explores the relationships between the use of agentive verbs (e.g., "selects", "organizes"), nominalizations (e.g., "development"), modal verbs (e.g., "may", "can"), copular verbs (e.g., "to be"), verbs of possession (e.g., "have"), and negation at the phrasal level (e.g., "not connected to the topic") and word level (e.g., "incomplete"), among other linguistic features of the performance descriptors in the rating scales. The present research focuses on the representation of writer or textual agency, as represented by agentive verbs, e.g., "skillfully manages paragraphing", and on the language of ability, as represented through the modal verbs *can* and *cannot*.

This study considers which of these linguistic features cluster together in depictions of writer agency and ability, how these linguistic features differ across

score or band levels, and the extent to which these different linguistic features represent the language of TOEFL versus IELTS performance descriptors. Ultimately, the implications of these linguistic choices are considered, including the significance of foregrounding rater judgment, writer agency, writer (in)ability, and/or seemingly static nominalized qualities of an essay.

3.2. FrameNet and the analysis of linguistic agency

To explore issues of writer and textual agency, even in the absence of overt agents (as is frequently the case in performance descriptors headed by verbs), this study employs the concept of semantic frames. In an experimental study on the relationship between agentive language and perceived blame, Fausey & Boroditsky (2010: 644) consider "agentive and nonagentive frames," defining a "canonical agentive description" as one in which there is "a person as the subject in a transitive expression describing a change of state". They further define a "canonical nonagentive description" as one which is "intransitive and does not place the person as the subject for the change-of-state event". The present study builds upon the basic distinction between "agentive and nonagentive frames," using FrameNet to capture finer distinctions in the semantic frames associated with specific verbs.

FrameNet is a database of lexical items and semantic frames, based on the theory of Frame Semantics (Fillmore 1976; Fillmore & Baker 2010). This theory suggests that word meanings can be derived in part by the semantic frames that are evoked by their use. For instance, the verb "kick" evokes the notion of an agent and a patient, someone who kicks and someone or something that is kicked: these thematic roles are the semantic frame associated with the verb "kick". Besides simple agent and patient roles, FrameNet also includes a number of other relevant roles. For instance, FrameNet also includes thematic roles like Goal, i.e., an agent's objective, such as demonstrating writing proficiency, and Cognizer, i.e., a person or entity that judges or perceives – in the case of rubrics, a rater. FrameNet is used in this study to explore how frequently agents or other roles are associated with the particular verbs featured in the TOEFL and IELTS rating scales.

While broader distinctions, such as agentive and nonagentive, are useful in experimental studies, such as Fausey & Boroditsky (2010), and corpus studies, such as Dryer (2013), the ability to draw more subtle distinctions is one of the chief contributions of a discourse analytic approach to this area of inquiry. Thus, the purpose of using FrameNet in this study is two-fold. First, it supplies a wider range of semantic frames associated with verbs; and secondly, it provides an independent judgment of the frames associated with the verbs in the rating scales.

3.3 The rating scales

The analysis focuses on the following rubrics: (1) the TOEFL iBT ® Independent Writing Rubrics, (2) the TOEFL iBT ® Integrated Writing Rubrics, (3) the IELTS Task 1 Band Descriptors (Public Version), and (4) IELTS Task 2 Band Descriptors (Public Version). A brief comparison with the SAT Scoring Guide for essays is also provided. All of these scales and level descriptions are publicly available through the official web sites for each of the exams, and may thus be employed by test-takers and institutions in test preparations and admissions decisions.

It is important to note, however, that these different rating scales demonstrate different orientations, as described by Alderson (1991). One significant distinction between the TOEFL and IELTS scales is that the public versions of the IELTS scales, analyzed in the present study, are examples of Alderson's user-oriented scales in that they are presented with the expressed purpose of providing information to test-takers, teachers, and administrators. They are not directly used in rating but instead in educating users regarding the criteria that are evaluated in the writing sections of the IELTS exam. The actual IELTS scales used for operational rating are not provided, though Cotton & Wilson (2011) provide some discussion of raters' response to actual IELTS writing descriptors for coherence and cohesion, ultimately suggesting "the possible need to fine tune some of the wording in the band descriptors" (2011: 52) in this area to improve construct validity. This additional information about the writing descriptors is clearly intended for a scholarly audience. In contrast, the TOEFL scales are both assessor-oriented and user-oriented in that the same scales function in two ways: these scales are used by actual TOEFL raters in assigning scores (Knoch et al. 2014: 60), and they are provided publicly online for educational and informational purposes. The SAT writing rubric similarly appears to be user and assessor-oriented, with no separate public version of its rubrics as well as indication on its web page that the standards provided there are the ones used in assessment.

4. Results

4.1. Agentive verbs in the TOEFL writing rubrics

The use of agentive verbs in the TOEFL rubrics is limited, but they are more frequent for the highest scores in both the Independent Writing Rubrics and the Integrated Writing Rubrics. For the Integrated Writing Rubric, descriptors are full sentences (as opposed to verb phrases or noun phrases). Contrasting the introductory sentences – and in particular the verb phrases or verbal forms in italics – for each score clearly illustrates the differences in terms of linguistic agency across descriptors for different scores, as seen in (1) below:

(1) a. **Score 5**: "A response at this level successfully *selects* the important information from the lecture and coherently and accurately *presents* this information…[ ]"
b. **Score 4**: "A response at this level *is generally good in selecting* the important information from the lecture and in *coherently and accurately presenting* this information in relation to the relevant information in the reading, but it may *have* minor omission, inaccuracy, vagueness, or imprecision…[ ]"
c. **Score 3**: "A response at this level *contains* some important information from the lecture and *conveys* some relevant connection to the reading, but it *is marked* by one or more of the following:"
d. **Score 2**: "A response at this level *contains* some relevant information from the lecture, but *is marked* by significant language difficulties or by significant omission….[ ]
e. **Score 1**: "A response at this level *is marked* by one or more of the following:"
f. **Score 0**: "A response at this level merely *copies* sentences from the reading, *rejects* the topic or *is otherwise not connected* to the topic…[ ]"

As can be seen in (1) above, the description for Score 5, with two agentive verbs "selects" and "presents", features the most linguistic agency. This description is followed by a pattern of copular verbs, verbs of possession, and passive voice dominating the descriptions at the lower levels.[3] The verbs "selects" and "presents" from the Score 5 descriptors manifest as gerunds in the Level 4 description with a copular verb followed by a predicate adjective: "is generally good in selecting the important information from the lecture and in coherently and accurately presenting this information." Already some of the active nature of these verbs is lost in favor of the copular verb and gerunds. The description for Score 3 shares syntactic features with those of descriptions at higher levels, i.e., an agentive verb "conveys"; similar levels, i.e., a verb of possession "contains"; and lower levels, i.e., passive voice "is marked by". The fact that this is a middle category is mirrored by the fact it shares syntax with the descriptors above and below it. The description for Score 2 loses the single agentive verb present in the Score 3 descriptions but maintains the verb of possession "contains" and the passive voice "is marked by". The Score 1 description features only the passive voice, i.e., "is marked by"; and then active verbs return in the description for 0, e.g., "merely copies."

---

[3] The descriptors for Score 0 are a seeming exception to this pattern, although they do not feature any of the same verbs as the descriptors for higher levels, thus not being truly comparable.

Comparatively, each level in the Independent Writing Rubric is framed by a brief introductory statement followed by a list of descriptors that are either verb phrases or noun phrases. The same basic pattern is still evident, even given this different format. For example, the Level 5 introductory statement (in bold) and descriptors (in the bulleted list) appear as follows, in (2) below:

(2) Score 5: "An essay at this level largely accomplishes all of the following:
- Effectively addresses the topic and task
- Is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details
- Displays unity, progression and coherence
- Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors"

Levels 4 and 5 are framed in terms of achievement with the agentive verb "accomplishes," as seen in the Level 5 descriptor in (2).According to FrameNet, the verb "accomplishes" invokes a frame in which an "Agent has been working on a Goal; the Agent manages to attain it".

While the Level 4 and 5 descriptors are introduced and framed by an agentive verb, i.e., "accomplishes", none of the verbs in the bulleted Level 4 or 5 descriptors for the Independent Writing rubric is canonically agentive. For instance, the verb "addresses," which actually describes what essays do with regard to the topic and task in Levels 3 through 5, may be best understood with the following frame: "A stretch of linguistic discourse or a Text that a Communicator produces has a Topic that it is about". For this frame, FrameNet lists examples like the following:

(3) "Ostrovsky addresses monetary policy in Chapter 5"

(4) "This book is mostly about particle physics"

While this frame can invoke simply the sense of what a piece of writing "is about", in instances in the TOEFL descriptors, a writer who "addresses the topic and task" does seem to feature some of the properties of an agent. He or she is sentient and volitionally writes on the given topic; whether the writing is perceived by the rater to be on-topic, however, is not completely within a writer's control. Thus it is ultimately the rater's decision of whether verb phrases like "addresses the task" apply to the writer/writing, which contrasts with the situation for many agentive verbs, e.g., "kick", where the conditions under which this verb would apply are

independent of a third party's judgment. The semantic frame for verbs of providing evidence, like "display", may even feature a role for a Cognizer, "The person for whom the Proposition is supported by the Support". In the context of a rating scale, many of the frames for such verbs have an implicit Cognizer, the rater, who ultimately judges whether a given descriptor applies. Still, relative to the lowest levels, Levels Four and Five are clearly described in more active terms, with many more action verbs, as can be seen in (2) above.

Action verbs in this rubric correlate with descriptors of higher levels of proficiency. Just as in the level introductory statements, action verbs are featured more frequently in the bulleted Level 4 and Level 5 descriptors. Notably, the mitigations provided in the dependent clauses at these levels, e.g., "though it may have minor lexical or grammatical errors", do not feature active voice action verbs. Instead, these mitigations feature verbs of possession, i.e. "have", "contain", and passive voice, e.g., "may not be fully elaborated". Flaws are chiefly something that Level 4 and 5 essays *have*; positive features are the result of something that Level 4 and 5 essays *do*.

In contrast to the agentive verb that introduces the Level 4 and 5 essay, passive voice introduces the Level 3 essay descriptors: a Level 3 essay "is marked by one or more of the following", followed by the list of descriptors. In this passive construction, the qualities listed after the colon (e.g., "addresses the topic and task using somewhat developed explanations, exemplifications, and/or details") may loosely be seen as the grammatical agents of the verb "mark", with the essay being the patient. In this sense, these qualities mark, or distinguish, a Level 3 essay. Alternatively, a rater may "mark" a Level 3 essay by identifying the listed characteristics. In either circumstance, this framing does not allow textual agency for the Level 3 essay, as the agentive verb "accomplishes" does for the Level 4 and Level 5 essays.

The introductions to the Level 2 and Level 1 descriptors also feature non-agentive language, as seen in (5) and (6) below:

(5) "An essay at this level may reveal one or more of the following weaknesses" (Level 2)

(6) "An essay at this level is seriously flawed by one or more of the following weaknesses" (Level 1)

A non-agentive verb phrase, i.e., "reveal one or more of the following weaknesses", and passive voice, i.e., "is seriously flawed" are the main verbs of the Level 1 and 2 introductory statements.

4.2. Nominalizations in the TOEFL writing rubrics

A related issue is the marked use of nominalizations in the lowest score descriptors for the Independent Writing Rubrics. While the presence of agentive verbs like "accomplishes", "addresses", "selects", and "presents" in the highest score descriptors conveys the sense of an essay (or writer) that is more active and agentive, nominalizations have the inverse effect for the lowest score descriptors. The introductory statements for Level 3 through Level 5 all conclude with simply "the following", as exemplified in (2) above. This structure in Levels 3 through 5 introduces descriptors that are headed by verbs, also seen in (2); comparatively Level 1 and 2 introductory statements add a noun after "the following" – "the following weaknesses". This language sets up a list of descriptors that are all nouns, as seen in (7) below:

(7) Score 2: An essay at this level may reveal one or more of the following weaknesses:
- Limited development in response to the topic and task
- Inadequate organization or connection of ideas
- Inappropriate or insufficient exemplifications, explanations or details to support or illustrate generalizations in response to the task
- A noticeably inappropriate choice of words or word forms
- An accumulation of errors in sentence structure and/or usage

With nominalizations such as "development", "organization", "connection", "exemplifications", "explanations", "generalizations", and "accumulation" being the defining syntactic feature of these descriptors, Levels 2 and 1 are not syntactically parallel with the other levels. It is not that other levels do not include nominalizations in their descriptions. Comparing the descriptors for Score 5, in (2), with those for Score 2, in (7), it is apparent that some of the nominalized forms are shared by descriptors for the highest level and those for a much lower level, i.e., "exemplifications", "explanations", and "choice". However, the most conspicuous difference is the absence of verb forms in the descriptors combined with an increased frequency of nominalization. For instance, some of the descriptors that involved verbal morphology in (2) for Score 5, i.e., "well organized and well developed", are represented as nominal forms for Level 2 in (7) above, i.e., "limited development" and "inadequate organization". Further additional nominal forms appear in the Level 2 and 1 descriptors that do not appear in any form for other levels, for example "an accumulation of errors" and "questionable responsiveness".

Nominalization is less frequent in the Integrated Writing Rubrics, but the same pattern is present: nominalization is a syntactic feature that more closely

correlates with lower scores. The only nominalization present in the Score 5 description for the Integrated Writing Rubric is negated: essays of this score feature minor language errors that "do not result in inaccurate or imprecise *presentation* of content or *connections*". In other words, all of the language that describes Level 5 essays is free of nominalizations; the only nominalizations present here describe what essays of this score do not do, contrasting these higher scoring essays with those having lower scores. Comparatively, Level 2, 3, and 4 essay descriptors discuss "omission", "connection", "usage", and "expressions" in un-negated contexts, as seen in (8) below:[4]

(8) a. "The response contains language errors or *expressions* that largely obscure *connections* or meaning at key junctures or that would likely obscure understanding of key ideas for a reader not already familiar with the reading and the lecture" (Score 2)
b. "Some key points made in the lecture or the reading, or *connections* between the two, may be incomplete, inaccurate, or imprecise" (Score 3)

This use of nominalization eliminates agency from the descriptors for these levels and serves the function of reifying processes. Rather than descriptions that feature a writer (or "essay") that "responds to the task" or "develops ideas", the language of the rubric turns parts of the writing process into discrete things. As Fowler (1991: 80) argues, with nominalization, "processes and qualities assume the status of things: impersonal, inanimate, capable of being amassed and counted like capital, paraded like possessions". The main verbs in these introductions produce a similar effect in terms of eliminating agency. The verb "reveal" in the Level 2 introduction, i.e., "An essay at this level may reveal one or more of the following weaknesses:", fits best with FrameNet's sense of the word "reveal" that pertains to providing evidence. This word evokes the following frame: "The Support, a phenomenon or fact, lends support to a claim or proposed course of action, the Proposition", which again features no agent. Certainly, a writer would not deliberately "reveal a weakness" in writing during an exam. As "volitional involvement in the event or state" (Dowty 1991: 572) is a core, defining property of agents, neither the writer nor the text can be understood as agentive in this statement. Similarly, in the introductory statement for the Level 1 descriptors, passive voice is used again, i.e., "An essay at this level is seriously flawed by one or more of the following weaknesses:". As with the Level 3 introductory

[4] As with the Independent Writing Rubrics, the very lowest levels of the Integrated Writing Rubrics feature language that is not comparable in terms of content or structure to the descriptors for the other levels.

statement, the characteristics identified in the descriptors have an impact on the essay, rather than the essay or the writer *doing* anything.

4.3. Agentive verbs and nominalizations in the IELTS writing band descriptors

Compared to the TOEFL rubrics, the IELTS rubrics include significantly more agentive verbs, e.g., "uses", "manages", "organises". The overwhelming majority of descriptors for these rubrics are headed by verbs, as seen in (9):

(9) a. "fully addresses all parts of the task" (Band 9, Task Response)
b. "produces frequent error-free sentences" (Band 7, Grammatical Range and Accuracy)
c. "makes inadequate, inaccurate or overuse of cohesive devices" (Band 5, Coherence and Cohesion)
d. "uses only a very limited range of words and expressions with very limited control of word formation and/or spelling" (Band 3, Lexical Resources)

Overall, verb forms are maintained across band descriptors for different scores in these rubrics, with distinctions largely encoded by adverbs, negation, and clausal modifiers, as seen in (10):

(10) a. **Band 9:** "fully *addresses* all parts of the task"
b. **Band 8**: "sufficiently *addresses* all parts of the task"
c. **Band 7**: "*addresses* all parts of the task"
d. **Band 6**: "*addresses* all parts of the task although some parts may be more fully covered than others"
e. **Band 5**: "*addresses* the task only partially; the format may be inappropriate in places"
f. **Band 4**: "*responds* to the task only in a minimal way or the answer is tangential; the format may be inappropriate"
g. **Band 3**: "does not adequately *address* any part of the task"
h. **Band 2**: "barely *responds* to the task"
i. **Band 1**: "answer is completely unrelated to the task"

The few exceptions to this pattern are generally in lower band descriptions, e.g., (10i) above, and in descriptors for Lexical Resources and Grammatical Range and Accuracy. For example, for an essay in Band 3, "some structures are accurate but errors predominate, and punctuation is often faulty". In all of these cases, the descriptor is a full sentence rather than a verb phrase, again differing from the TOEFL rubrics, one of which features descriptors headed by nouns for

some of the lower scores. It is in these full sentence descriptors and clausal mitigations that one sees the decreasing level of linguistic agency in lower band descriptors. For instance, in (10d) we see a passive voice construction in the dependent clause, i.e., "although some parts may be more fully covered than others". In the descriptions for Bands 5, 4, and 1 in (10e, f, and i), we see a copular verb and predicate adjective: e.g., "the format may be inappropriate in places". All of these constructions feature non-agentive verbs or reduced agency resulting from passive voice. Thus some of the same basic pattern of reduced agency is apparent across band descriptors in the IELTS rubrics, if not to the same degree as seen with the TOEFL rubrics.

Though nominalization is not as much of a defining feature of the IELTS descriptors, the same basic pattern of increased nominalization for lower score descriptors is occasionally apparent in these rubrics, as well, as seen in (11):

(11) a. "logically *organises* information and ideas" (Band 7)
b. "presents information with some *organization*" (Band 5)

In this example, the verb "organises", which appears in the higher band descriptor instead appears as the nominalized form "organization" in the lower band descriptor.

Overall, however, the descriptors for Bands 3 through 9 are relatively syntactically parallel in terms of their use of agentive verbs and nominalizations in these rubrics. It is actually the use of modal verbs *can* and *cannot* in the lowest band descriptors that creates a larger difference both in the syntax and in the nature of the descriptors at these lowest levels in the IELTS rubrics.

4.4. Language of ability in the IELTS writing band descriptors

One of the distinguishing features of both the IELTS Task 1 Writing Band Descriptors and the Task 2 Writing Band Descriptors is the use of language denoting ability rather than actual performance. This linguistic strategy represents a departure from the language of other ESL and EFL writing rubrics, like the TOEFL rubrics, as well as rubrics that assume native-speaker test takers, like the SAT writing rubric. This language of ability appears exclusively in the lowest band descriptors, and is limited to descriptors for Lexical Resources and for Grammatical Range and Accuracy, as seen in (12) below:

(12) a. "cannot use sentence forms except in memorized phrases" (Band 2)
b. "cannot use sentence forms at all" (Band 1)
c. "can only use a few isolated words" (Band 1)

Comparatively, descriptors for these areas in Bands 3 through 5, which also discuss an essay's limitations in the same areas, employ references to the reader, the writer's "attempts", or qualified statements of successful performance, as seen in (13):

(13) a. "errors can cause some difficulty for the reader" (Band 5)
b. "uses only basic vocabulary which may be used repetitively or which may be inappropriate for the task" (Band 4)
c. "attempts sentence forms but errors in grammar and punctuation predominate and distort the meaning" (Band 3)

Even descriptors for Task Achievement and Coherence and Cohesion for the lowest bands in these rubrics, which feature almost no agentive language, still avoid language of (in)ability, instead opting for copular verbs, verbs of possession, and action verbs that are negated or otherwise mitigated, as seen in (14):

(14) a. "barely responds to the task" (Band 2)
b. "answer is completely unrelated to the task" (Band 1)
c. "has very little control of organizational features" (Band 2)

This is clearly an area in which the performance descriptors for the different bands and even within the bands differ syntactically. It is of course possible to frame these other descriptors in terms of ability or inability. For instance, the descriptor "fails to communicate any message" (Band 1) could just as easily be expressed as "cannot communicate any message". It is interesting to note the level of confidence these descriptors suggest in their assessment of the test taker's inability to use English grammar or lexical resources, even while similar judgments are withheld for Task Response and Coherence and Cohesion. However, there is a distinction between performance and ability. That this distinction is collapsed for Bands 1 and 2 may reveal a bias in the language of these rubrics.

4.5. Key similarities and differences between descriptors in the TOEFL and IELTS rubrics

In summary, the TOEFL and IELTS rubrics share a pattern of discourse that comments on writer agency and/or ability. They also both feature variation between the descriptors for the highest and lowest scoring essays in the extent to which descriptors employ agentive language and/or the language of ability. However, there are important differences between the language of the TOEFL writing rubrics and the language of the IELTS writing rubrics and how they comment on writer agency and/or ability. Linguistically, a higher concentration

of agentive verbs for descriptors associated with the highest scores and a higher concentration of nominalizations in descriptors for lower scores characterize the TOEFL writing rubrics. Similar linguistic patterns are presented in Dryer (2013) for college classroom-based rubrics. Comparatively, the IELTS rubrics feature generally more agentive language across band descriptors for different levels. The key feature distinguishing the descriptors for the lowest bands is the presence of language of ability. Explicit language of ability is absent from the TOEFL rubrics.

4.6. Linguistic agency and the language of ability in the SAT scoring guide for essays

It may be argued that the differences in linguistic agency and the language of ability in score and band descriptors of different levels merely reflect the real variation in writers' performances at different levels. Under this view, the use of more agentive verbs in higher band and score descriptors and the use of more nominalizations and language of (in)ability in the lowest score descriptions neatly corresponds with the higher degree of agency that the most skilled and proficient writers demonstrate and conversely the lesser degree of control of language and rhetorical moves demonstrated by less proficient writers. Surely no one would doubt that more proficient writers enjoy more control of their language and rhetorical choices, but writers at all levels are able to do *something* with words. Recall for instance the language of the CEFR descriptors, which were very deliberately "worded in positive terms, even for lower levels" (North 2007: 657).

In short, there are distinctions between a writer's performance and the language that test designers, raters, or educators use to describe that performance, just as there are distinctions between performance and ability. Perhaps this is not more evident than in the descriptors for another large-scale writing assessment's rubric: the SAT Scoring Guide for writing. Compared to the IELTS and TOEFL rubrics, the SAT Scoring Guide features verb syntax that is relatively more parallel, with descriptors at different levels featuring similar levels of linguistic agency and consistently avoiding language of ability. When one compares the descriptors for a score of 6, in (15), and those for a score of 1, in (16), the similarities in language across descriptors of different levels becomes apparent:

(15) Score of 6
An essay in this category *demonstrates* clear and consistent mastery, although it *may have* a few minor errors. A typical essay:
- Effectively and insightfully *develops* a point of view on the issue and *demonstrates* outstanding critical thinking, using clearly appropriate examples, reasons and other evidence to support its position

- *Is well organized and clearly focused*, demonstrating clear coherence and smooth progression of ideas
- *Exhibits* skillful use of language, using a varied, accurate and apt vocabulary
- *Demonstrates* meaningful variety in sentence structure
- *Is free of most errors* in grammar, usage and mechanics

(16) Score of 1
An essay in this category *demonstrates* very little or no mastery, and *is severely flawed* by ONE OR MORE of the following weaknesses:
- *Develops* no viable point of view on the issue, or *provides* little or no evidence to support its position
- *Is disorganized or unfocused*, resulting in a disjointed or incoherent essay
- *Displays* fundamental errors in vocabulary
- *Demonstrates* severe flaws in sentence structure
- *Contains pervasive errors* in grammar, usage or mechanics that persistently interfere with meaning

Without glossing over the marked differences between essays scoring 6 and those scoring 1, these descriptors preserve some syntactic similarities across scores. First, both sets of descriptors are framed in terms of what the essay "demonstrates". The frame for this verb, according to FrameNet, can either be related to causing a perception, i.e., "An Agent, Actor, Entity, or Medium causes a Phenomenon to be perceived by a Perceiver", or providing evidence, with Support of a Proposition supplied to a Cognizer who interprets the evidence. While the first sense of the word allows more agency for the text/writer, both senses recognize the presence of a third party who judges or perceives the evidence. This word choice is not specifically different from the choice of "reveal" in a similar statement in the TOEFL rubrics, i.e., "An essay at this level may reveal one or more of the following weaknesses" (Independent Writing Rubrics, Score 2). However, the TOEFL rubrics are not consistent in employing this type of language. For instance, the semantic frame of evidence and perceiver introduces the entire set of descriptors for Score 2 on the Independent Writing Rubrics, as seen above, while verb frames that do not assume a perceiver (i.e., the rater) appear at higher levels, e.g., "accomplishes". The consistent use of a semantic frame pertaining to evidence and perceivers in the SAT Scoring Guide descriptors also contrasts with the language of ability used for lower band descriptors in the IETLS rubrics, where performance in a particular domain is not discussed in terms of evidence or perception of a reader.

The single biggest contrast between the descriptors for Score 6 and Score 1 in the SAT Scoring Guide mirrors a similar syntactic distinction between descriptors for different scores in the TOEFL rubric. According to the SAT Scoring Guide, a high scoring text "may *have* a few minor errors" while a very low scoring text "*is severely flawed* by ONE OR MORE of the following weaknesses:". Again, as with the TOEFL rubrics, higher scoring essays "have" errors, while lower scoring essays are characterized by the passive voice construction "is flawed by". Beyond this difference, the rest of the verb syntax and semantic frames are the same.

The individual descriptors feature remarkably similar verb syntax, especially around areas that correlate with linguistic agency and the language of ability. The Score 6 essay "develops" and "demonstrates" its points, while the Score 1 essay descriptor for this same area employs the verbs "develops" and "provides". The organization and focus of the Score 6 essay is discussed with a copular verb followed by past participles, i.e., "is well organized and clearly focused". Comparatively, these same features in the Score 1 essay are also described with the equivalent copular verb and participles: it "is poorly organized and/or focused". While the Score 6 essay "exhibits skillful use of language," the Score 1 essay "displays fundamental errors in vocabulary", both featuring verbs of evidence. The verbs of evidence continue in the descriptions of sentence variety, where the Score 6 essay "demonstrates meaningful variety" while the Score 1 essay "demonstrates severe flaws". Finally where the Score 6 essay "is free of most errors," the Score 1 essay "contains pervasive errors," both being described in terms of features that are contained (or not) in the document.

Compared to the descriptors in the TOEFL and IELTS rubrics, the SAT writing performance descriptors are framed relatively more similarly across different scores. Differences in the SAT descriptors are achieved by negation, modifiers, and noun complements rather than in the verb syntax and semantic frames. This is significant because this strategy avoids the unnecessary framing of certain essays as *doing* more, with more agentive verbs, while others are characterized by a list of nominalized and static qualities, as in the TOEFL rubrics. It also avoids the unnecessary conflation of performance and ability in the IELTS rubrics. Finally in its choice of verbs, it systematically and across score levels recognizes the presence of a reader whose perceptions ultimately inform the rating of a given essay.

5. Discussion and conclusion

There is variation in the language of agency and ability in large-scale second language writing assessment rubrics, both between rubrics and between different levels' descriptors. But "the distinction between language ability and

the performance of that ability [is] at the same time a central axiom and a dilemma for language testing" (Bachman 1990: 308).Bachman goes on to contrast "real-life" (RL) and "interactional authenticity" (IA) approaches, explaining that "[o]ne problem with the RL approach, then, is that it treats the behavioral manifestation of an ability as the trait itself. Language tests, however, like all mental measures, are indirect indicators of the abilities in which we are interested" (1990: 309). Comparatively, "the IA approach views authenticity as residing in the interaction between the test taker, the test task, and the testing context…Both the development and selection of authentic language tests is thus based on a theoretical framework that includes the language abilities of the test taker and the characteristics of the testing context" (1990: 322). Messick (1988: 5), in his discussion of validity, puts similar emphasis on the importance of context for the interpretation of scores and their generalizability. Though Davies (2008) and Fox (2007) suggest that the IELTS exam is informed by Bachman's IA approach, IELTS band descriptors for lower scores do feature language that describes ability rather than performance or "behavioral manifestations of an ability". The language of ability is absent from all of the writing rubrics for large-scale admissions exams, such as the SAT, ACT, and GRE; it is also absent from the TOEFL rubrics. Thus its appearance in the IELTS rubrics is somewhat remarkable.

Variation in depictions of linguistic agency across band descriptors creates a similar issue: in ascribing more or less agency to essays that are scored at different levels, the rubric (and test makers) collapse the distinction between ability and performance. Though some of the differences in agentive language across band descriptors, e.g., the increased frequency of nominalizations in descriptors for lower scores, may be intended to soften a negative critique, this syntactic variation creates other issues. In particular, the language of the rubrics asks readers to make a judgment about the amount of agency and ability that writers have, rather than to describe their actual performances in syntactically parallel terms. This same pattern was observed in Dryer (2013) for rubrics that assume writers who are native speakers, but it is not an unavoidable feature of writing rubrics, as seen with the SAT rubric, which features descriptors that are more syntactically parallel in terms of verb syntax. Descriptors that focus on actual performance rather than assumed writer agency or ability may avoid some of the problem outlined by Bachman (1990) of conflating performance and ability.

The language of performance descriptors continues to be an item of interest for language testing and writing assessment, but one that to this point has not received much dedicated scholarly treatment, especially concerning the use of verbs, nominalization, and linguistic agency. Instead, the topic often emerges in the discussion of other questions on language testing and writing assessment. For instance, in a recent volume of *Language Testing,* a special issue on

"Assessing oral and written L2 performance: Raters' decisions, rating procedures and rating scales", Kuiken & Vedder (2014: 282) summarize some of the concerns and questions about rating scales raised in the issue:

> Another theme discussed by all authors in this issue concerns the use of rating scales. The following questions were addressed: What makes a good rating scale? Can this scale be employed for different tasks, with different learners, in different contexts and across different target languages? […]

A more detailed awareness of the syntax of performance descriptors may help shed light on these questions. For instance, one fruitful area for future research might be empirical investigation of the extent to which the variation in linguistic agency and the language of ability described in this study affects perceptions of actual writer agency or ability for different audiences, including raters, teachers, students, and administrators. If many scholars agree that using degree modifiers alone is insufficient to distinguish levels, perhaps an awareness of other areas of syntax could also inform performance descriptors or have an impact on the effectiveness of a rating scale.

In considering whether a rating scale may be used in different contexts, for different tasks, or with different audiences, we might also consider how learners and educational institutions understand the language of these scales. In addition to their large-scale national and international use, the TOEFL and IELTS rating scales are used or adapted by schools for their own language assessments (Becker 2010: 126), and the language of these descriptors impacts programs' recommended scores for admission (Golder et al. 2011). While not trained as raters – students, teachers, and schools are the intended audience of these publicly available rating scales; and in the absence of training, different audiences often have different interpretations of the language of rating scales (Li & Lindsey 2015), which sometimes proves vague or underspecified even for trained raters.

Schaefer (2008) was actually interested in how raters with less training, who would approach the essays and rating process more as "lay readers" (2008: 471), might respond to L2 writing and the rating process. In an effort to explore rater bias in the assessment of EFL essays, Schaefer (2008) analyzes the ratings of a sample of essays written by Japanese learners of English by 40 native English speaker raters. Though all of the raters in this study were English teachers in Japan and received some brief training on rating, Schaefer deliberately selected less experienced raters for the purpose of exploring "which features of EFL writing, as operationalized in the rating scale categories, would influence NES raters' severity and leniency in their judgment of essay quality" (2008: 471). He concludes "that raters are more likely to have severe or lenient bias towards higher ability writers than lower ability writers. Extremely low

ability writers also tend to attract more bias interactions". While higher and lower ability writers attracted more bias interactions than other writers, the direction of the bias depended to some extent on individual rater idiosyncrasies: some raters rated higher ability writers more leniently, while others rated them more severely. The same was true of essays representing the lowest performance (2008: 486, 489).

One would certainly not expect the rater variability and bias found in Schaefer's study in large-scale assessments, which involve more extensive rater training; however, there is a similar pattern of variability in the language of performance descriptors for such large-scale assessments. As seen in the analysis and discussion above, the performance descriptors for the writing assessment portions of the IELTS and TOEFL exams feature different syntax, especially around the language of ability and agency, for the very highest and lowest bands. In the IELTS rubrics, language of ability appears only in the very lowest band descriptors; and while active verbs – including agentive verbs – dominate in the IELTS rubrics, the single place an active verb is conspicuously absent in the highest band descriptors is with regard to errors. So much agency is awarded to the highest scoring essays that even when these writers do make errors, the descriptors do not characterize the writers as making errors; instead "rare minor errors occur only as 'slips'" in Band 9 essays. Comparatively, essays representing Bands 5, 6, 7, and 8 all "make" errors. In Bands 3 and 4 errors "predominate"; and finally essays representing Bands 1 and 2 "cannot use sentence forms". In the TOEFL rubrics, linguistic agency is stripped from the performance descriptors for the lowest scoring essays, which feature almost exclusively nominalization, while the highest score descriptors feature the most agentive verbs on the rubric.

In other words, in these rubrics, the syntax of the descriptors used to characterize the very highest and lowest scoring essays differs, and in ways that correlate with one pattern of bias found in Schaefer (2008): more lenient treatment of higher scoring essays and more severe treatment of lower scoring essays. The IELTS Band 9 descriptors give the benefit of the doubt to the high scoring essay, distancing the writer from any errors by removing the active verb phrase "make errors" and instead simply indicating the presence of errors that "occur". The language of the lowest band descriptors judges these essays comparatively severely: the language of these descriptors is confident enough to remark on the (in)ability of the writers rather than their performance. Far from giving the benefit of the doubt, this language actually characterizes the lower scoring essay in more severe terms than what might be warranted by the evidence provided by a single exam.

Further research might shed light on how raters react to descriptors featuring more or less linguistic agency and language of ability. However, since the

rubrics analyzed for this study are all pubic versions, perhaps an equally important question is how teachers, students, and administrators (the actual audience for these particular public rubrics) respond to this language. Dryer (2013) explores the significance of varying language across performance descriptors in college-level writing rubrics. Making reference more generally to the brief and often negated language of lower level performance descriptors, he argues the following:

> The impoverished language in the lower traits is ironic in this context, since the rhetoric of absence and negation that operate on those levels do little to scaffold teachers' understanding of the causes and complexities of writing appraised at those levels or to provide opportunities for these readers and writers to recognize themselves as agents able to do things differently next time (2013: 27).

While the context for the rubrics in Dryer's corpus is different from that represented by a large-scale language assessment, like the TOEFL or IELTS exams, there are still some similarities in terms of applied significance.

Since the TOEFL and IELTS rubrics are publicly available and used as reference, assessment, and educational tools by administrators, teachers, and students, it is important to consider how the language of these rubrics might inform teachers' and administrators' views of these students and the students' views of themselves. The current language of these rubrics may be unnecessarily disempowering for students whose essays are rated with lower scores. Variation in the level of linguistic agency and the language of ability across descriptors for different bands or scores is not a necessary feature of writing rubrics, and it may indeed reflect bias for or against the most or least proficient writers. However, Messick (1988: 117) argues that "the validation of test use should assure that adverse social consequences do not stem from any source of test invalidity". As seen in previous studies of linguistic agency and perception (e.g., Fausey & Boroditsky 2010, Fausey et al. 2010, and Chakroff et al. 2015), less agentive language causes perceptions of less actual agency. Thus Dryer's concern over the impact of syntactic variation in writing rubrics might be well-founded, potentially playing out on a larger scale for the TOEFL and IELTS exams. The field may benefit from more focused research on the language we use to describe and assess student performance (and ability). Critical attention to the linguistic details of our rubrics can only serve to reduce potential biases in the language of assessment and to empower students across all levels of proficiency.

REFERENCES

Ahearn, Laura M. 2001. Language and agency. *Annual Review of Anthropology* 30. 109–137. DOI: 10.1146/annurev.anthro.30.1.109

Alderson, J. Charles. 1991. Bands and scores. In J. Charles Alderson & Brian North (eds.),*Language testing in the 1990s: The communicative legacy*, 71–86. London: Modern English Publications/British Council.

Alderson, J. Charles, Neus Figueras, Henk Kuijper, Guenter Nold, Sauli Takala & Claire Tardieu. 2004. *The development of specifications for item development and classification within the Common European Framework of Reference for Languages: Learning, teaching, assessment: Reading and listening: Final report of the Dutch CEF Construct Project*. Lancaster University. http://eprints.lancs.ac.uk/44/1/final_report.pdf.

Bachman, Lyle F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Banerjee, Jayanti, Xun Yan, Mark Chapman & Heather Elliott. 2015. Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing* 26. 5–19. DOI: 10.1016/j.asw.2015.07.001

Becker, Anthony. 2010. Examining rubrics used to measure writing performance in US intensive English programs. *The CATESOL Journal* 22(1). 113–130.

Billig, Michael. 2008. The language of critical discourse analysis: The case of nominalization. *Discourse & Society* 19(6). 783–800. DOI: 10.1177/0957926508095894

Brindley, Geoff. 1998. Describing language development? Rating scales and SLA. In Lyle F. Bachman & Andrew D. Cohen (eds.), *Interfaces between second language acquisition and language testing research*, 112–140. New York: Cambridge University Press.

Bucholtz, Mary & Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies* 7(4–5). 585–614. DOI: 10.1177/1461445605054407

Calkins, Lucy McCormick. 1994. *The art of teaching writing* (new ed.). Portsmouth: Heinemann.

Chakroff, Aleksandr, Kyle A. Thomas, Omar S. Haque & Liane Young. 2015. An indecent proposal: The dual functions of indirect speech. *Cognitive Science* 39(1). 199–211. DOI: 10.1111/cogs.12145

Cotton, Fiona & Kate Wilson. 2011. An investigation of examiner rating of coherence and cohesion in the IELTS Academic Writing Task 2. https://www.ielts.org/-/media/research-reports/ielts_rr_volume12_report6.ashx

Covill, Amy E. 2012. College students' use of a writing rubric: Effect on quality of writing, self-efficacy, and writing practices. *Journal of Writing Assessment* 5(1). http://journalofwritingassessment.org/article.php?article=60

Crusan, Deborah. 2010. *Assessment in the second language writing classroom*. Ann Arbor: University of Michigan Press.

Davies, Alan. 2008. *Assessing academic English. Testing English proficiency 1950–89: The IELTS solution*. Cambridge: Cambridge University Press.

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67(3). 547–619.

Dryer, Dylan. 2013. Scaling writing ability: A corpus-driven inquiry. *Written Communication* 30(1). 3–35. DOI: 10.1177/0741088312466992

Duranti, Alessandro. 2004. Agency in language. In Alessandro Duranti (ed.), *A companion to linguistic anthropology*, 451–473. Malden, MA: Blackwell. DOI: 10.1002/9780470996522.ch20

Ehrlich, Susan. 2001. *Representing rape: Language and sexual consent*. New York: Routledge.
Fausey,Caitlin M. & Lera Boroditsky. 2010. Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review* 17(5). 644–650. DOI: 10.3758/PBR.17.5.644
Fausey, Caitlin M., Bria L. Long, Aya Inamori & Lera Boroditsky. 2010. Constructing agency: The role of language. *Frontiers in Psychology* 1. 162. DOI: 10.3389/fpsyg.2010.00162
Fillmore, Charles J. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences* 280. 20–32. DOI: 10.1111/j.1749-6632.1976.tb25467.x
Fillmore, Charles J. & Collin Baker. 2010. A frames approach to semantic analysis. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 313–339. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199544004.013.0013
Fowler, Roger, Bob Hodge, Günther Kress & Tony Trew. 1979. *Language and control*. London: Routledge.
Fowler, Roger. 1991. *Language in the news: Discourse and ideology in the press*. London: Routledge.
Fox, Janna D. 2007. *Language testing reconsidered*. Ottawa: University of Ottawa Press.
Golder, Katherine, Kenneth Reeder & Sarah Fleming. 2012. Determination of appropriate IELTS Writing and Speaking Band Scores for admission into two programs at a Canadian post-secondary polytechnic institution. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée* 14(1). 222–250.
Hambleton, Ronald K.& Mary Pitoniak. 2006. Setting performance standards. In Robert L. Brennan (ed.), *Educational measurement* (4th ed.), 433–470.Westport, CT: Praeger.
Hawkey, Roger & Fiona Barker. 2004. Developing a common scale for the assessment of writing. *Assessing Writing* 9(2). 122–159. DOI: 10.1016/j.asw.2004.06.001
Henley, Nancy M., Michelle Miller & Jo Anne Beazley. 1995. Syntax, semantics, and sexual violence: Agency and the passive voice. *Journal of Language and Social Psychology* 14(1–2). 60–84. DOI: 10.1177/0261927X95141004
Jeffery, Jill V. 2009. Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing* 14(1). 3–24. DOI: 10.1016/j.asw.2008.12.002
Knoch, Ute. 2007. *'Little coherence, considerable strain for reader'*: A comparison between two rating scales for the assessment of coherence. *Assessing Writing* 12(2). 108–128. DOI: 10.1016/j.asw.2007.07.002
Knoch, Ute. 2009. Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing* 26(2). 275–304. DOI: 10.1177/0265532208101008
Knoch, Ute. 2011. Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing* 16(2). 81–96. DOI: 10.1016/j.asw.2011.02.003
Knoch, Ute, Susy Macqueen & Sally O'Hagan. 2014. An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT® writing test. *ETS Research Report Series* 2014(2). DOI: 10.1002/ets2.12038
Kuiken, Folkert & Ineke Vedder. 2014. Raters' decisions, rating procedures and rating scales. *Language Testing*. 31(3). 279–284. DOI: 10.1177/0265532214526179
LaFrance, Marianne & Eugene Hahn. 1994. The disappearing agent: Gender stereotypes, interpersonal verbs and implicit causality. In Camille Roman, Suzanne Juhasz & Cristianne Miller (eds.), *The women and language debate: A sourcebook*, 348–362. New Brunswick, NJ: Rutgers University Press.

Li, Jinrong & Peggy Lindsey. 2015. Understanding variations between student and teacher application of rubrics. *Assessing Writing* 26. 67–79. DOI: 10.1016/j.asw.2015.07.003

Lumley, Tom. 2002. Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing* 19(3). 246–276. DOI:10.1191/0265532202lt230oa

Matsuda, Paul Kei, & Jill V. Jeffery. 2012. Voice in student essays. In Ken Hyland & Carmen Sancho Guinda (eds.), *Stance and voice in written academic genres*, 151–165. New York: Palgrave Macmillan. DOI: 10.1057/9781137030825_10

Messick, Samuel. 1988. Meaning and values in test validation: The science and ethics of assessment. *ETS Research Report Series* 1988(2). DOI:10.1002/j.2330-8516.1988.tb00303.x

Morales, Meghan Corella & Jin Sook Lee. 2015. Stories of assessment: Spanish–English bilingual children's agency and interactional competence in oral language assessments. *Linguistics and Education* 29. 32–45. DOI: 10.1016/j.linged.2014.10.008

North, Brian. 2007. The CEFR illustrative descriptor scales. *The Modern Language Journal* 91(4). 656–659. DOI: 10.1111/j.1540-4781.2007.00627_3.x

North, Brian& Günther Schneider. 1998. Scaling descriptors for language proficiency scales. *Language Testing* 15(2). 217–262.DOI: 10.1177/026553229801500204

[OED =] *Oxford English Dictionary* (3rd ed.). Oxford: Oxford University Press. https://www.oed.com/

Schaefer, Edward. 2008. Rater bias patterns in an EFL writing assessment. *Language Testing* 25(4). 465–493. DOI: 10.1177/0265532208094273

Spandel, Vicki. 2006. In defense of rubrics. *English Journal* 96(1). 19–22.

Upshur, John A. & Carolyn E. Turner. 1995. Constructing rating scales for second language tests. *ELT Journal*49(1). 3–12. DOI: 10.1093/elt/49.1.3

Winke, Paula & Hyojung Lim. 2015. ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing* 25. 38–54. DOI: 10.1016/j.asw.2015.05.002

Wodak, Ruth & Michael Meyer (eds.). 2009. *Methods of critical discourse analysis* (2nd ed.). Los Angeles, CA: Sage.