Social Contexts of Writing Assessment: Toward an Ecological Construct of the Rater

Dylan B. Dryer and Irvin Peckham

Abstract

Research on rater behaviors has historically interpreted raters' decision-making practices as decontextualized. This article suggests raters' scores are better interpreted as a residue of the overlapping social systems in which raters are enmeshed: historical currents in the field of writing studies, the day-to-day dynamics of the room within which the assessment occurs, the ongoing micro-ecology of scoring at their table, and near-instantaneous intrapersonal responses to all of the above. We argue that the use-validity of assessment data is bolstered by a fuller understanding of the relationships behind the complex social behaviors of raters, and we make four recommendations to bring writing-assessment practices into closer alignment with contemporary complex models of the writing construct. The analysis and recommendations are based on triangulated research focused on a single assessment (19 readers; 152 first-year composition portfolios; 370 sets of scores) using a Phase 2 Portfolio assessment protocol.

Although no assessment can capture the full complexities or predict the complete range of a student's writing abilities (Condon), assessment scholars agree that writing assessments based on a specificity about a particular domain of locally valued writing skills (Huot), careful scoring procedures (Lane and Stone), and development of criteria (Broad et al.; Hambleton and Pitoniak) can provide useful information about students' abilities to meet clearly defined writing outcomes.[1] The most current interpretations of validity also demand attention to an assessment's social consequences (Huot, O'Neill and Moore; Kane; Poe and Inoue), e.g., the risk of misdirecting students to non-credit, writing courses with significantly higher rates of attrition and failure than mainstream credit-bearing alternatives or

the disparate impact of a poorly designed rating scale on federally protected populations.

In response to growing awareness of the social costs of construct-underrepresented and construct-invalid assessments, the field has begun to call for models that better approximate the complexity of students' actual writing practices and development (Wardle and Roozen). It is also time, we contend, to acknowledge the complexity of raters' reading practices (Barkaoui; Harsch and Martin; Knoch "Rating"). Though large-scale assessment technologies such as rater calibration and rubrics are rhetorically constructed as translocal—i.e., inevitable, transparent, and ideologically neutral (Huot and Neal)—in the final analysis, each writing assessment is an irreproducible confluence of idiosyncratic humans and complex textual artifacts in specific spaces and times (Elliot). Researchers, however, must not shrink from engaging the complexities of local factors; any cost incurred in protecting students and teachers from the consequences of overgeneralizations on the basis of numbers derived from an always-partial assessment is worth paying (Moffett).

Inattention to local factors also helps explain why the state of knowledge about raters' decision-making practices remains so fragmented. Although Liz Hamp-Lyons observed nearly twenty years ago that a "great deal remains unknown about . . . raters' rating processes" ("Rating" 761), fifteen years of research advanced the field's understanding very little, if her recent comments on the subject are any indication ("Writing Assessment" 4; see also Zhang). Meanwhile, the variety of construct-irrelevant influences on raters' scoring decisions seems endless: assumptions about writers' skill levels (Diederich), prior scoring experiences (Barritt, Stock and Clark; Vaughan; Wolfe "Relationship"; Wolfe "Uncovering"; Wolfe, Kao and Ranney), prompt fatigue (Weigle), handwriting (Powers et al.), personality type (Callahan; Lumley and McNamara), candidates' choice of genre (Carrell) or prompt (Hamp-Lyons and Mathias), severity drift (Myford and Wolfe), scoring order (Singer and LeMahieu), inferred gender (Haswell and Tedesco Haswell), and ego involvement with the projected writer (Dryer "Mirror"; Wiseman). Raters score in perplexing ways for curious reasons, and their "ontological and epistemological orientation" usually remain unknown quantities (Elliot, Briller, and Joshi 6). To paraphrase Charles Bazerman, in every assessment situation, students' papers disappear into the "black boxes" of raters' nervous systems (Wolfe and McVay 5).

It is curious that raters' minds should remain black boxes—after all, composition studies reoriented to a socio-cognitive view of communication a quarter-century ago. While writing-assessment tools and practices have generally been sluggish to reflect these advances (Behizadeh and Engel-

hard; Dryer "Scaling"), if the field has accepted the utility and accuracy of socio-cognitive models for curricular purposes, these models should also be extended to writing assessment. To that end, this empirical, qualitative research study attempts to pry open the black box by resituating raters' decisions in a complex ecology of scoring. To avoid the flattening effects of single-method inquiries (see Suto), this study joins other recent mixed-method approaches (e.g., Knoch "Investigating") by employing MacMillan's model of Ecological Inquiry. This model allows us to see raters' decisions and the possible social consequences of the scores assigned as the residue of the operations of four social contexts in a particular assessment: 1) *Field* level effects, which are the raters' tacit and explicit beliefs about what writing is and how it should be taught, organizational systems (e.g., program, university, or corporation), pre-assessment teaching conditions, status differences among raters; 2) *Room* level effects, which are the organization and purpose of the particular assessment; 3) *Table* level effects, which are the social ecology of each table of raters; and 4) *Rater* level effects, which are raters' cognitive and affective reactions to *Field*, *Room*, and *Table* effects.[2]

Methods

The assessment scene we studied as participant-ethnographers was sponsored by Jennings-Baker,[3] publisher of a composition textbook called *The College Rhetor*.[4] We learned of the assessment, part of a Jennings-Baker initiative called PASS,[5] from a JB representative, who gave us permission to study the assessment. We offered our own time as raters in exchange for time appropriated from paid raters for interviews. As provisional members of the research project, we were included in communications with the principals as details for the reading were finalized, and along with the other seventeen raters, we went through the training and calibration sessions over two days of scoring.[6] In figure 1, we offer a floor plan of the assessment: at Table A in the top right are sitting the table leader (TL-A) and the four raters: A1, A2, A3, and A4. Table B has three raters working with their table leader (TL-B), and so on for Tables C and D.

We supplemented our notes and observations with data from video/tape-recordings of the assessment and semi-structured interviews (Prior 187) with raters from different tables during both days of the assessment. Interview questions were conducted after the training and before the assessment (addressing motivations for participating, impression of training, and questions of value), midway through the assessment (addressing impressions of and factors in their scoring decisions), and post-assessment (addressing comparison of this scoring experience to other scoring experiences, overall

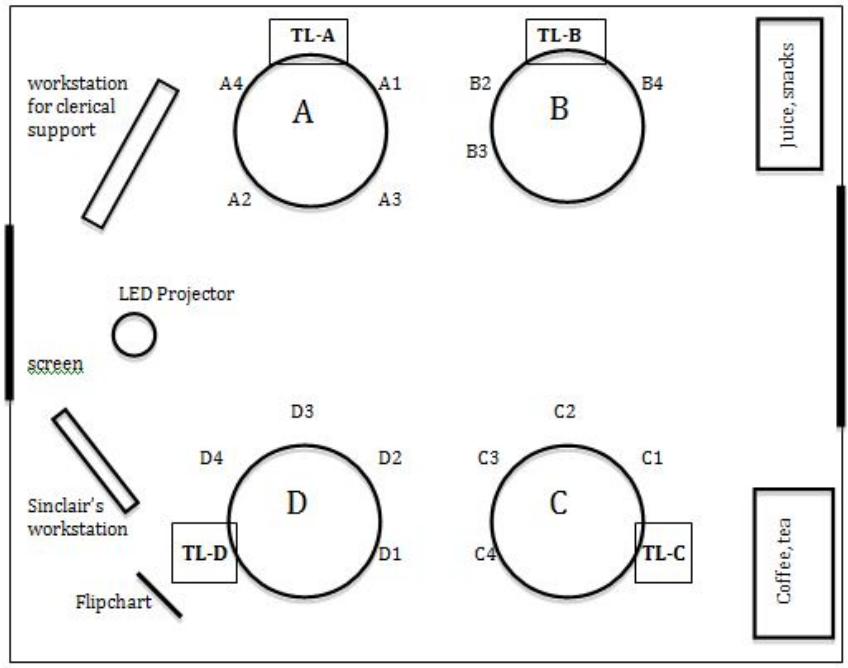


Fig. 1. Floor Plan of PASS Assessment

descriptions of their decision-making processes, and unanticipated issues in the scoring experience). We conducted an exit survey of all raters immediately after the assessment ended, as well as focus-group discussions with the four table leaders and five of the other raters an hour later. Finally, we conducted a quantitative analysis of the holistic scores and of the four analytic scores, which were based on the WPA Outcomes Statement 1.0: Rhetorical Knowledge; Critical Thinking, Reading, and Writing; Writing Processes; and Knowledge of Conventions.[7]

Site Context

JB retained Michelle Sinclair,[8] a national figure in US writing assessment, as chief reader. She explained to raters on the first day that the PASS assessment had three purposes: 1) to establish a protocol for publisher-teacher joint assessments of the effectiveness of textbooks generally; 2) to determine the effectiveness of the *College Rhetor* specifically; and 3) to establish a benchmark against which teachers using other textbooks and other outcomes-based assessments could measure their own students' achievements.[9] Sinclair and JB selected raters from teachers in two- and four-year colleges who had used the *College Rhetor* and were familiar with the WPA Out-

comes. Incentives included an hourly rate of about twenty dollars, staying at the resort hotel where the assessment was held, and the opportunity to learn more about assessment. The two of us were assigned rater numbers D4 and A4 and joined seventeen other raters for the two days of scoring.

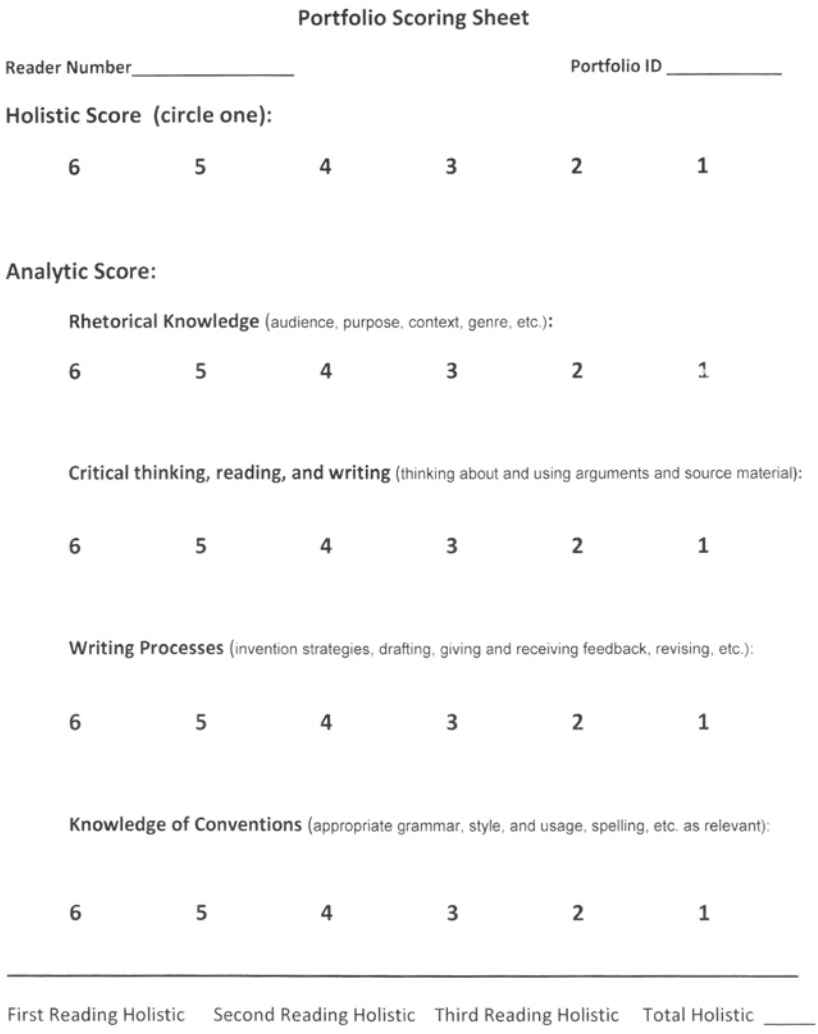
Raters were calibrated using the "Phase 2" portfolio assessment model (White). Raters were briefed on White's argument that the general impression valued in holistic scoring loses instrumental reliability when raters consult different documents by the same student; portfolio assessments should therefore focus on students' reflective introductions, consulting the contents only "to authenticate what the student is saying in the reflective letter" (592-4). In addition, Sinclair reminded readers that this assessment was low-stakes (Baker). "Nothing we do here will affect any student," Sinclair said. "You're not grading, you're producing data."

The scoring sheet used in this assessment, seen in figure 2, is modeled directly on the WPA Outcomes Statement 1.0 and required raters to use 6-point scales to assign a holistic score and four analytic scores. Raters were to total their analytic scores—a sum that could range from 4 (each outcome ranked a 1) to 24 (each outcome ranked a 6). Portfolios were randomized and scoring was adjudicated—on any portfolio, the two raters' holistic scores had to be at least adjacent, and the sums of their two total trait scores had to be within four points. If either or both of these criteria were not met, the portfolio was referred to a third rater, usually one of the four table leaders.

Analysis of Social Systems Discoverable in the PASS Assessment

In what follows, we examine raters' production of scores within the complex ecology of this assessment, illuminating quantitatively derived scoring patterns with information derived from qualitative methods. To provide a glimpse of this complexity, we add another layer to figure 1 to represent how raters' decisions are influenced by the entire ecology of the assessment scene, not just by procedures intended to influence them (such as calibration).

To keep the figure visually accessible, only ecological effects on Table C are mapped. Beginning with rater C4, who is sitting to the left of his table leader and with his back to rater D1, we know that he is more than meets the chief reader's eye. He claimed to have enjoyed the handful of assessments he'd participated in prior to his invitation to read for the PASS project and felt well prepared by the calibration sessions. He was less sure,



Portfolio Scoring Sheet

Reader Number_____ Portfolio ID _____

Holistic Score (circle one):

6 5 4 3 2 1

Analytic Score:

Rhetorical Knowledge (audience, purpose, context, genre, etc.):

6 5 4 3 2 1

Critical thinking, reading, and writing (thinking about and using arguments and source material):

6 5 4 3 2 1

Writing Processes (invention strategies, drafting, giving and receiving feedback, revising, etc.):

6 5 4 3 2 1

Knowledge of Conventions (appropriate grammar, style, and usage, spelling, etc. as relevant):

6 5 4 3 2 1

First Reading Holistic Second Reading Holistic Third Reading Holistic Total Holistic _____

Fig. 2. PASS Scoring Sheet

however, how closely PASS priorities matched his own. He was unsure whether he scored like his tablemates; while he often wondered how they were scoring, he never consulted his table leader and was lukewarm about the importance of the scoring sheet, believing strongly that he scored best from his initial impression. "I read this stuff all the time," he explained in an interview, "and it's often the case that your first-paragraph impression will be confirmed by the rest"—a common assumption among relatively unpracticed raters.

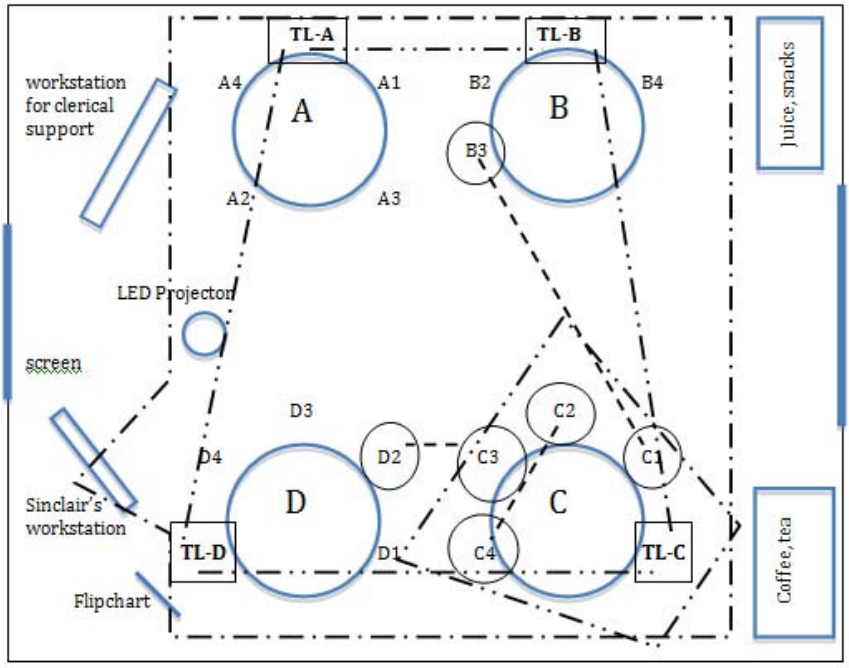


Fig. 3. Floor Plan of PASS Assessment with Sample Effects on Table 2's Scoring

The circles around the raters at Table C indicate their own sense of professionalism and priorities—as well as vested interests in managing their own labor—which are only partially susceptible to official influences, here symbolized by the encompassing dash-dotted line that emanates from the chief reader's station and that bounds all four tables.

Although raters typically read in small groups around tables, we know little about these microecologies of scoring (but see Colombini and McBride). Yet these ecologies exert hidden influences on scoring behaviors, illustrated here by dashed lines. For example, rater C4 was, in addition to the profile he divulged above, also an irritant to rater C2.

> C2: At our table, we have someone who's a little more vocal while he's scoring, so I was kind of influenced by him at the very get-go. I knew which portfolio he was referring to. I found myself saying, 'I disagree and let me find all the reasons I disagree.' . . . I ended up scoring that paper lower than the rest of the room . . . I wonder whether I was just overreacting to his praise of it, you know.
>
> Peckham: It's like if *he* said that, it can't be that?

> C2: He said it was really good, and if he hadn't said that, I would have given it a 4, but for some reason I gave it a 3.

Such influences extend beyond the imagined community of "our table" or "we table leaders" (mapped here with dot-dot-dashed lines): for example, the effect of rater D2's squeaky chair on rater C3's nerves; the preference of rater C1, who explained that she was "from a community college," to leave the calibration discussion to those "university instructors in there" after rater B3 "misinterpreted my comment about fluency" during the initial calibration session.

In our analysis, we document a sampling of the effects we found operating at each of the four levels described above: field, room, table, and rater. We will argue that field-, table-, and rater-level data is essential information about raters' reading practices and should enrich assessment constructs that remain oversimplified by the tendency to report only the final, room-level data. In other words, we are arguing for a vision of assessment congruent with contemporary constructs of writing.

Social System of Field

A JB representative introduced Sinclair to the raters as a national leader in writing assessment research—a kind of incarnation of the field. As chief reader, Sinclair had to quickly establish the ethos of the scoring sheet and the expertise of the raters. Sinclair argued that the WPA Outcomes Statement (WPA OS) represented a field-wide consensus on achievable aims of first-year composition courses. She then used the scoring sheet's (see fig. 2) alignment with the WPA OS to establish room-level confidence by characterizing it as "a professional scoring guide that depends on your expertise as a teacher." In the next twenty minutes, Sinclair reiterated "professionalism" three more times: that the data produced by the study required "professional judgments" of the raters; that raters should not "score like a computer; [but] like a professional, which you are"; and that raters' appraisal of explicit or implicit learning in the reflective letter would necessitate "professional judgment." These compliments were justified: most raters were experienced teachers and raters. Rater A3, the least experienced teacher in the room, had five years' teaching experience; nearly half had more than twenty, and the rest at least ten. Experience with assessments was more uneven—for three raters, this was their first controlled assessment experience. All raters, however, considered themselves qualified to cope with rating issues raised in the calibration session. On our survey, the lowest self-

assigned ranking of familiarity with issues raised in the calibration sessions was 4 on a 6-point scale.

To capture the raters' sense of the field, one of our exit survey questions was to "assign a rank ordering" to the four traits that appeared on the PASS scoring sheet. In the following discussion of their rankings, we operate with the assumption that that the ordering of outcomes in the WPA OS reflects the relative importance of each outcome to the Outcomes Collective, the name assigned to the large group of rhetoric and composition scholars who collaborated over several years to author the Statement (Harrington *xvi*). The WPA OS had to appear in some order, obviously, but that order was not arbitrary and reflected the scholarly consensus of the field—the social-epistemic over the current-traditional model (Berlin). Peckham, who was one of five members of the steering committee responsible for finalizing the 2000 version of the WPA OS, remembers that in both the 1998 Chicago meeting and the discussions of the steering committee, the ordering of the WPA Outcomes Statement was a conscious, rhetorical construction—as one would expect from rhetoricians. Since the scoring sheet was adopted from the WPA OS, field-level effects should be observable in the relative distribution of overturned scores—for example, by raters who claimed to value Critical Thinking, Reading, and Writing as most important compared to those who claimed to value Conventions most highly.

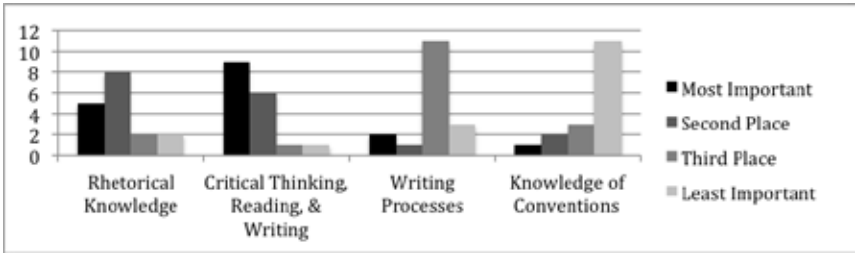


Fig. 4. Relative Priority Given to each WPA Outcomes Statement Trait

In figure 4, the four WPA Outcomes are shown with the relative importance each of the seventeen raters assigned them. Raters broadly reproduced the Outcome Statement's ordering.[10] The two outcomes Rhetorical Knowledge and Critical Thinking, Reading, and Writing in the majority of raters' perceptions were more important than Processes and Conventions. Since raters identified themselves by rater number on the exit survey, we were able to link each rater's ranking of the WPA OS traits to their individual scoring history, represented in data table 1.

Table 1 Percentage of Overturned Scores by Rankings of Outcomes

Ranking of Outcome	**Percentage of Overturned Scores**
Ranked Rhetorical Knowledge Most Important	16.6
Ranked Rhetorical Knowledge Second	18.1
Ranked Rhetorical Knowledge Third	18.1
Ranked Rhetorical Knowledge Least Important	17.7
Ranked Critical Thinking Most Important	13.0
Ranked Critical Thinking Second	16.0
Ranked Critical Thinking Third	30.5
Ranked Critical Thinking Least Important	49.6
Ranked Process Most Important	42.2
Ranked Process Second	18.8
Ranked Process Third	13.3
Ranked Process Least Important	16.8
Ranked Conventions Most Important	30.0
Ranked Conventions Second	20.0
Ranked Conventions Third	22.3
Ranked Conventions Least Important	13.1

This linkage of ranking to scoring history enables us to see the distributions of overturned scores vis-á-vis individual raters' valuation of outcomes. The distribution of overturned scores does not precisely replicate the ordering in the WPA OS, since the raters prioritized Critical Thinking over Rhetorical Knowledge. Those who ranked Critical Thinking first had only 13% of their total scores overturned; the rater who ranked Critical Thinking least important had 49.6% of his scores overturned. Process and Conventions appear in their familiar third and fourth places, respectively, and those who placed them there were unlikely to have their scores overturned. Raters who elevated either Process or Conventions above its station, so to speak, were much more likely to be overruled. The lack of any significant correlation between how raters ranked Rhetorical Knowledge and the percentage of overturned scores is curious; it's possible that in the current zeitgeist, it has been displaced by Critical Thinking as a proxy for curricular aims (Peckham *Going* 105-11, but see also Maid and D'Angelo).

Just as every score is an incremental addition to a rater's ongoing development of her beliefs about what writing is and how it should be appraised,

every assessment makes its own contribution to this conversation at field-level social systems. Mapping raters' self-positioning with respect to field-level articulations of principles against their decision-making practices could offer an important source of information about how deeply those field-level articulations have penetrated into local training and program practices as well as offering ground-level glimpses of how these principles might be shifting.

Social System of Room

The room is the tacit arbitrator of appropriate score values. Sinclair explained that "the scoring should depend on the criteria we've agreed on, rather than on who scores it." As chief reader, Sinclair had to gain raters' confidence while persuading them to accept scores on benchmark portfolios and room-determined scores as the preferred score. After raters scored the benchmark portfolios for the first calibration session, table leaders led discussions before the scores were tallied and shown on a flipchart. In her subsequent discussion, Sinclair affirmed the room's "professional judgment" by uniting references to the room (underlined) with references to herself and the team of readers who chose the benchmark portfolios (in **boldface**):

> It's really interesting to see the spread for the first paper. We'll come back to that, but let's see where the areas of most agreement are. And this is pretty clear. Ten out of 16 scorers scored it a 1. That's what **we** saw it as. [Pointing to an isolated high score] This person is obviously way out of line, whoever it is; I don't know, and I don't care, but whoever that person is, you should re-evaluate how you're proceeding because the rest of the room essentially said '1 or at best a 2' OK? Let's see where else we are in agreement. Paper B, 12 out of 17 said it was a 5; **we** saw it as a 5, and again, these people are sort of close, and these 2 people are out of line. By definition, right? The room saw it as a 5. We're going to go back and talk about these in a minute; let's just see how it works . . . Paper C, now, the room has it as a 3, but **we** saw it as a 2, and we'll need to talk about whether it could be better there. Paper D, clearly the best of the lot, and the room clearly saw it as a 6. [points to a low score] This person is coming from some other planet; [light laughter] maybe this was a clerical error. But that's a very strange score for a paper that everybody—almost everybody—thought was the best of the lot. Portfolio E **we** saw as a 4; the room saw it as a kind of 4-3. We'll talk about what's the better score.

Such rhetorical framing of portfolio quality as having an ontological existence is routine in writing assessment, as it was during both days of the

PASS assessment. Drawing raters' attention to the areas of greatest consensus around a 2 and a 5 during the second calibration session, Sinclair suggested that "I think we're entitled to say that [Portfolio] L *really is* a 5, and [Portfolio] M *really is* a 2, in the sense that 'really' is the group consensus. So you want to check your score against those two" (emphases added).

Several raters claimed in our interviews to have shifted their scoring in response to room-level calibration. Rater D2, for example, initially scored the benchmark 6 portfolio as a 4, privately believing that the portfolio's sophisticated language and syntax was mostly, as she put it, "BS." During the discussion, however, the room successfully rationalized the portfolio's assertions about knowledge of syntax and conventions as a kind of rhetorical knowledge—the writer's astute anticipation of how teacher-readers would be reading his or her text, an anticipation that in fact turned out to be the case.

> Rater D2: I am aware now that I don't necessarily have the same values as the others *in the room*, so I have tried to be more liberal and reward articulate writing, even though there may not be much substance to it *in my mind*. When I assign the scores right now, I realize that if people have the conventions down, that's being valued, and so I am being more generous. People *in the room* are interpreting that as [the student's] 'rhetorical knowledge.' (emphases added)

Rater A3's distribution of scores replicated the room's with a maximum difference at score point 3 (see data table 3) because he has been able to internalize the room norms. He explained that he made difficult decisions by bearing in mind that

> inasmuch as I don't have access to prompts, course goals, or the students who wrote these papers . . . the knowledge that this is not a grade helps to overcome that. Grades play so much a part of motivating students or rewarding things that maybe don't show up in a final draft or things like that. Since this isn't a grade in that sense, it kind of helps mediate that complication.

Similarly, Rater C2 disciplined herself against overreacting to what she perceived to be second-language interference:

> I hear [Sinclair's] voice in the back of my head as I look at this terrible grammar. . . . I'm not sure whether I'm giving them a genuine score or I'm deflating them because of her direction. . . . I'd give it a 1, but I guess I'll give it a 3 or a 4 because she said not to be harsh on them.

While both raters A3 and C2 may initially read the portfolios through their previous experiences, the values of the room, as a social system, mediates their initial response.

It was inevitable, however, that some raters' sense of the field as discussed would produce some interpretations of the "right" score that would be at odds with the room-level social system Sinclair was attempting to construct. Rater D1 explained that she "liked the idea of structuring the assessment around the WPA Outcomes," (and was among those who ranked Critical Thinking, Reading, and Writing first) but complained that the scoring sheet didn't allow her to evaluate a student's voice. Her distribution drifted lenient (see data table 3 below) in comparison to the room level distribution. Others who clung to private scoring schema drifted severe: rater B3, feeling that a lot of the portfolios she read "were kind of in the middle," made decisions by eliding critical thinking with a private outcome of engagement:

> Rater B3: When there's some real thought, that's more of an upper-level paper, even if they're not the greatest writer. There's more critical thinking, there's better analysis, there's more sense of personal responsibility in their learning. It isn't the teacher's job to make you learn in college, it's your responsibility.
>
> Peckham: You'd say that the interactive element between the teacher and the class actually has something to do with the score you assign? That's not in the criteria, is it?
>
> Rater B3: I definitely value engagement. I definitely valued students who don't blame me for their lack of learning.

Rater B3's private appraisal of how the students might have interacted with a projected version of herself (note the shift from past to present tense: "I . . . valued" to "students who don't") is neither supported by the scoring sheet nor measurable in the portfolios. Typical of raters assessing a construct-underrepresented trait, she trends severe, awarding nine overturned 3s. As we will explain further below, we are not suggesting that voice or engagement cannot or should not be assessed, but we know that raters reduce scoring criteria to "manageable representation[s]" to limit the demanding cognitive routines necessary to appraise a complex construct such as writing ability (Bejar 4). Raters D1 and B3 perhaps did not have sufficient opportunities to translate their commitment to voice and engagement into the room-level language drawn from the scoring sheet.

Social System of Table

The sound of raters reconciling (or not) private motives with room-level intentions will have been familiar to most readers, especially those who have followed the scholarly uptake of Bob Broad's landmark study of an assessment scene at City College nearly fifteen years ago. But by shifting our focus from the social system of the room to the table, we can see how different configurations of personalities responded differently to Sinclair's attempt to socialize raters/tables into similar scoring behaviors. In data table 2, we have constructed a thumbprint for each table, measured against the thumbprint of the room. We might describe Table B as timid, assigning a disproportionate number of scores to the middle, the well-known 3-4 category and reluctant to risk low and high evaluations (see Elliot 103). Table D, by contrast, leads the pack toward a flattened score distribution pattern, assigning 12% 1s, 13% 2s, 19% 5s, and 13% 6s. Raters at this table were bold, unafraid of ranking portfolios as marginal or exceptional.

Such scoring thumbprints can be instructive to compare, not least because they can illustrate variance concealed in a room-level mean. Table-level scoring thumbprints can also indicate the effects of an elusive influence on scoring behaviors: the table leader. Straddling room and table systems (Hoskens and Wilson), table leaders play a pivotal role in how individual raters score, as can be seen in how two different table leaders attempt to broker Sinclair's version of the field to the configuration of raters at their table.

Table 2 Distributions of Scores—Room v. Tables

	Table A	**Table B**	**Room Mean**	**Table C**	**Table D**
Score of 1	4%	1%	7%	10%	12%
Score of 2	15%	12%	14%	16%	13%
Score of 3	22%	28%	22%	19%	21%
Score of 4	31%	34%	28%	24%	23%
Score of 5	22%	19%	21%	23%	19%
Score of 6	7%	5%	8%	8%	13%

As we might expect from the distribution of scores at Table D, their table leader was a highly experienced rater and table leader. We can see her confidence and poise as a table leader in this transcript of her conversation with rater D2, whom she suspects has been scoring high:

> TL-D: This is the second paper I saw that you gave a 5 to; you gave it a 5, and I gave it a 3—and here's something I find interesting.
>
> D2: [with mild sarcasm] Interesting?
>
> TL-D: I think you were reading more into it. . . .
>
> D2: That's interesting. That's possible.
>
> TL-D: I think you were giving it a lot more benefit of the doubt . . . The best way we can proceed is to continue to refer back to the scoring guide. . . .
>
> D2: Actually, you're liberating me because I had a sense that if I *were to really go with my true feelings, the score would be a lot lower. . . .* [The benchmark 3 portfolio] just threw me. That paper was atrocious.
>
> TL-D: Even if you think it was atrocious, you have to look at it as a benchmark for *what we mean by a 3*. (Emphases added).

TL-D deftly negotiates the line between maintaining table sociability and ensuring that members of her subsystem are aligned with the room's macrosystem. If she senses resentment in D2's laughing echo of "Interesting?," she responds by guiding the conversation away from abstractions about "atrociousness" and back to local dynamics and the shared artifact of the scoring sheet. Doing so allows D2 to save face with some graciousness of her own saying, remarkably, that the table leader's intervention has in effect liberated her from her idiosyncratic sense of the field.

Table leader A, however, struggled to keep her table attuned to room-level dynamics, as we can see in this exchange with rater A1, who in this excerpt had just told her curtly that she "didn't care" whether or not the table leaders were checking on scoring.

> TL-A: You know what [calibration] is going to show us? This is not necessarily going to show us that we're wrong, but that if we're way off, maybe we need to take more time, and if we're not, the way we're doing is working. You see what I'm saying?
>
> A1: Another thing it can tell you of course is just what bothers you the most; I mean there are things that wouldn't significantly distract me that significantly distract *you*.
>
> TL-A: That's right—personal prejudices, and I understand that [pause] um I'm a uh grammarian, big time. But that doesn't distract me. . . .

> A1: It depends which it is. I can get by if they don't put a comma before [inaudible] but if you see some random comma and it's so distracting that it messes up the rest of the sentence.
>
> TL-A: Yeah, I know. Plus we're so skilled at putting in what we think it should be even though we see that it's not that; that sometimes in a way can get you through a paper [pause]. Well, we've all done it; same experience regardless of where we teach [trails off]

Unlike her colleague at Table D, table leader A fails to resist the interpersonal realm, electing to try to repair relations with A1 at the cost of losing focus on a room-level issue (i.e., the relative importance of conventions). By invoking off-rubric criteria, she attempts to be professional by invoking a shared teaching experience rather than by translating Sinclair's room-level calibration to the table. Perhaps as a consequence of this idiosyncratic uptake on conventions, table A generated twice as many nonadjacent conventions scores as Table D, as well as displaying a marked central tendency in scoring.

Raters react differently to ecological influences—conceding to some room-level imperatives, reserving resistance for others, negotiating prior allegiances to previous training and value-systems with the imperatives of room-level reliability. While so far we have seen raters' responses to "official" sources of rater influence (the chief reader and her table leaders), this conversation captured at Table A suggests that official sources are only part of the spectrum of influences. Here, rater A2 tries to open a discussion about the difficulty of assessing rhetorical knowledge, but is interrupted by A1, whose intervention appears to have contributed to Table A's asymmetrical relationship with the room:

> A2: If I'm on the fence, though, if they say they're addressing a particular audience, but I can't see one—
>
> A1: [*interrupting*] —at some point, I just give myself a break: 'would you pass this paper or wouldn't you pass it?' And that's your gut [snaps fingers] after reading an intro—gut, right? *Then* you kind of go through and justify.

Although Sinclair had warned the room "that you're going to have to fight against your tendency to do here what you've done before," A1 elected to limit her own labor by using a private grading scale. Unchecked by her table leader, this rater-level decision becomes a table-level factor in scoring. On the second day of scoring, A1 again interrupted the table leader to argue for a rating system used by piecework raters whose primary motive is to stay employed by standardized testing agencies.

> A2: What if I have like, hypothetically, a reflective [reflection] that doesn't reference the portfolio?
>
> TL-A: Sometimes what [the student] is expressing we might not recognize as one of the outcomes, but it actually is, in a roundabout way. And you go back into their papers and see do they have this? Whether it's addressed or not, that's when you go into the analytic scores and—
>
> A1: [interrupting]—that's where you get into that portfolio thing, though, 'cause it's only a 1-point differential. Who's gonna give it a 4? Nobody. Even if you give it a 2, everybody's going to give it a 1 or a 3; you're both going to be within a point.
>
> A2:—yeah [pause] yeah, I mean, I scored the SATs, so yeah.

We find evidence for the influence of this conversation in data table 3, which reports seat-level distributions relative to that of the room, identifying leniency and severity drift where measurable. As can be seen, A1 gave no 1s, preferring the safe card of the 2. In research comparing timed with untimed essay performances, Peckham finds the same logic governed piecemeal readers for the ACT Essay exams, a notable reluctance to award 1s and 6s on a 6-point scale. Rather, raters were clearly deciding whether a paper was a 2/3 or 4/5, as if knowing that second raters would also edge toward the middle ("Online" 130). Like those ACT raters, A1 awarded the highest proportion of 4s in the room; by the end of scoring, A2 had also drifted toward the middle, over-assigning 2s and 3s and under-assigning 5s.

Since raters could of course stay within a "typical" distribution relative to the room and still award a nonadjacent score, data table 3 also reports the number of overturned scores each rater produced, which helps us locate effects of even quite subtle inter-rater influences. "If we are to be totally honest for the sake of your data," reported C2, "there is one person who keeps kind of irritating me . . . because of hopping on the iPhone all the time and checking on the email all the time." Resentful that the table leader appeared to be asking everyone to pick up this rater's slack, rater C2 said she felt like saying, "No, I don't want to do her work. I want to check my email now. . . . I'm annoyed with her." C2 believed that she scored leniently, writing on her exit survey that the assessment procedure "influenced me to be more compassionate than I could have been before." Resentful of her off-task tablemate and irritated by the chatty rater C4 (as mentioned above and who awarded elven overturned 2s), C2 has lost touch with both the room and herself, over-scoring 1s and 2s and under-scoring 4s.

Table 3 Seat Distribution of Scores (N =1840)

	Score of 1	Score of 2	Score of 3	Score of 4	Score of 5	Score of 6
Room Distribution (overturned scores)	6.7% (40)	14.1% (46)	22% (55)	28% (43)	20.8% (56)	8.4% (19)
Table A: 465 scores						
Table Leader	0%	12%	36%†	24%	21%	6%
Rater A1	0%	8%	22%	45%* (6)	20% (9)	6% (5)
Rater A2	6%	23%† (1)	32%† (6)	24% (5)	10%† (2)	6%
Rater A3	8%	14% (3)	19% (6)	27% (6)	23% (3)	10%
Rater A4	1%	19%	10%*	30% (2)	33%* (2)	7%
Table B: 410 scores						
Table Leader	1%	24%† (5)	23%	30%	19%	3%†
Rater B1[1]	—	—	—	—	—	—
Rater B2	4%	11% (2)	26% (7)	40%* (1)	17%	2%†
Rater B3	0%	10% (1)	39%†† (9)	30% (1)	13%† (2)	9%
Rater B4	0%	7% (1)	25% (3)	36% (2)	25%	7%
Table C: 521 scores						
Table Leader	4%	18% (1)	28% (1)	22% (1)	22% (2)	6%
Rater C1	11%	9%	14%*	25% (3)	34%** (6)	8%
Rater C2	14% (3)	22%† (4)	22% (6)	12%	18%	12% (1)
Rater C3	14% (5)	6%*	6%** (2)	37%* (4)	27%* (10)	11% (3)
Rater C4	6% (1)	21% (11)	21% (5)	26% (5)	19% (5)	6% (1)
Table D: 444 scores						
Table Leader	4%	10%	23% (1)	34%	19%	10%
Rater D1	3%	3%*	18% (1)	33% (2)	23% (5)	21%** (2)
Rater D2	0%	4%*	22% (3)	29% (1)	26% (4)	18%* (5)
Rater D3	32%†† (27)	23%† (16)	22% (5)	8%†† (1)	13%† (4)	2%† (2)
Rater D4	8% (4)	28% (1)	16%	18% (3)	14%† (2)	18%*

Drift

* lenient—1 SD from room mean
**significant leniency—2 SD from room mean
† severe—1 SD from room mean
††Significant severity—2 SD from room mean

Note

1. Rater B1 did not attend the assessment session due to illness.

Raters were even influenced by raters they could not see. Rater A3, for example, was asked to adjudicate Portfolio 78, which B2 had scored a 5 and D3 (predictably, given his sharply severe scoring profile) had scored a 1. Not knowing D3's scoring profile, A3 struggled against his initial inclination to give Portfolio 78 a 5, ultimately lowering his score to as a way to accommodate D3's 1.

> RA3: I could have seen either a 5 or a 1 as a defensible position. Ultimately I gave it a 4; I thought it was worthy of being a higher half. But knowing that it got a 1, it encouraged me to look for reasons why it could have earned a 1.
>
> Dryer: So the presence of that other judgment [trails off]
>
> RA3: Absolutely.

In a 2011 review of neurophenomenological research, Marilyn Cooper describes all humans' attunement with influences in their immediate environment as structural coupling, a "process of mutual adaptation that occurs when organisms or systems perturb one another in a prolonged interaction, gradually becoming more attuned to one another" (437; see also Heath and Luff; for an assessment-specific example, see Columbini and McBride, esp. 202). We believe scoring tables are best similarly understood as permeable social systems, subject to multiple and competing influences. As we will see below, so are raters themselves.

Social System of Rater

Finally, we find effects in the intrapersonal realms of affect and cognitive dissonance. Rater B4 described her reluctant accommodation to the social system of the room, following the discussion of the benchmark 6 portfolio in the first calibration session:

> B4: If I have to give it a 5, I will. We had this conversation at our table yesterday; I wanted to see it the way we were asked to see it. There were those who said, 'where's the evidence in it?' and those who said, 'but look at the sentence structure!' The language was there; to me it was someone who could speak the language, but they weren't understanding what they were saying.

Rater B4 is willing to align but retains her private response to the portfolio. We can imagine that others engage in private resistance—perhaps muttering at the table or the buffet line, perhaps leaving a silent trail of nonadjacent scores. Nearly a third of rater D3's scores were 1s, as if he were waging a private war against the room's "voice of the majority." Others find ways to redress bias or perceptions of inadequate instruction in their scoring, developing workarounds to the room-level randomization design. Rater D2 told us she always guessed at the gender of the writer when reading the reflective essay. Although she worried that gendering caused her "to fill in the blanks for writers" based "on the space I think writers should occupy," she conceded that she would find "fluffy" writing more acceptable from a female than a male writer.

Some raters, recruited from nearby colleges, claimed that they could deduce the student's school from the reflective letter. As rater C2 explained: "While you're judging the paper . . . I'm also kind of judging the English department of that school." Asked what she meant, C2 explained, "When I know [the portfolio] comes from that more 'hippy' school, I'm a little more impressed with the student who came out of this casual environment but came out of it with an academic tone."

These comments complicate common assumptions about raters' behaviors. One might imagine that raters, as a consequence of calibration sessions, scoring guides, and prior experiences, base their scores on their perceptions of the portfolios in question measured against their interpretations of objectified standards. We found, to a surprising degree, raters reacting to and rejecting the scoring values of the room, table and field social systems. Interpreted within an ecology of scoring, a score is not the private appraisal of an isolated artifact but is instead produced by the dynamic interaction of multiple systems.

One of the more interesting findings in our study is raters' awareness of their own situation within these systems. Our entrance survey collected the sorts of information that are routinely used to assess rater expertise and compatibility (number of years' experience teaching writing, prior assessment experiences, current job description), and our exit survey posited a series of questions to which raters could respond with a ranking on a 6-point scale (e.g., familiarity with issues discussed during training and norming; preparation to assess the portfolios). We found that only the handful of questions that gave raters opportunities to reflect on their own scoring practices and table-dynamics consistently corresponded with nonadjacent scoring. Data table 4 reports those questions, mapped to the percentages of nonadjacent scores that each rater who responded in the top, middle, and bottom thirds of the scale produced. Rater D3, for example, rated his familiarity with the issues discussed during training and norming as 6 on a 6-point scale. However, when asked later in this same survey to indicate the degree to which those sessions changed the way he would otherwise have scored these portfolios, D3 circled 2, in the bottom third of the scale. Perhaps because he imagined himself as professional, a self-image Sinclair urged in the opening session, D3 assumed enough authority to resist table and room systems. He marked 6 in response to our question on frequency of thinking about how other raters were scoring but that didn't mean he was trying to score as they were scoring. He wrote in the margins of the survey, "Thought about it continually, but disregarded any and all *qualms*" (emphasis added)—a revealing way to characterize his relationship to his table, the room, and to assessment practices more generally. Placed in

Table 4 Selected Exit Survey Questions

Question	Identical/Adjacent Scores	Nonadjacent Scores
How closely did calibration fit with what you value in student writing?		
Not at all (1-2)	65.2%	34.8%
Fairly closely (3-4)	75.0%	25.0%
Very closely (5-6)	72.5%	27.6%
Do you feel that the other readers valued the same things in undergraduate writing as you?		
Very Little (1-2)	68.2%	31.8%
Somewhat (3-4)	74.3%	25.7%
Very Much (5-6)	72.6%	27.3%
Please indicate the importance of the scoring guide to your scoring.		
Not important (1-2)	71.4%	29.6%
Somewhat (3-4)	75.8%	24.2%
Very important (5-6)	79.4%	20.6%
How frequently did you know what score you would assign after reading the first two or three paragraphs of the reflection?		
Almost Always (5-6)	50.5%	49.5%
Sometimes (3-4)	65.2%	34.8%
Almost Never (1-2)	61.7%	38.3%
How often did you think about how the other readers or the table leader might have scored an essay when you assigned your score?		
Seldom (1-2)	51.2%	48.3%
Sometimes (3-4)	62.5%	37.5%
Often (5-6)	59.9%	40.1%

full ecological context, Rater D3 becomes more than just a seriously discrepant scorer, accounting for more than a fifth of all overturned scores. When his responses are mapped to his scoring profile, he offers a useful

insight into a kind of professional motivation that he surely shares with others.

Conclusion

"Without a clear understanding of . . . how scoring criteria must be resolved against an assessor's intuitive professional understanding of a piece of work," asks Victoria Crisp, "can we ever be really sure of what an assessment is measuring?" (10). Empirical qualitative research gives us insight into raters' understandings of the writing construct they are scoring and provides a potentially useful protocol to establish connections between scores and these understandings. Moreover, it suggests that a view of scoring from only the room level obscures important differences in table-level scoring distributions and rater-level reading dynamics. If raters' scores are residues of social dynamics of tables, the ethos projected by their table leader and chief reader, and by raters' complex relations to their working conditions in and concepts about the field, our exploration into the ecologies of an assessment site suggests four opportunities for working *with* local ecologies and conditions—not strategies to minimize or manage their influence.

First, research based on the ecology of scoring should inform scoring protocols. Raters' generally accurate perceptions of their own scoring profiles suggest that they have a useful part to play in identifying validity threats. Raters also deserve more opportunities to report and reflect on their perceptions alongside the conventional room-level calibration sessions, fraught as those sessions are with imbalances of authority, expertise, and assertiveness (see Rater C1's comment, above). Table leaders might also be encouraged to supplement chief readers' reliability reports with attentive interpretation of the activity at their table. Compare the reports of interpersonal conflict and tension we have documented at Table C with their table leader's characterization:

> TL-C: My people even without knowing me came to me a lot with 'You've got to look at this one, I don't know where to go.' When I've had experiences norming before, they've not been as nice . . . but everyone was like 'Hey, what do you think?'; 'Can I have your help with this?' People were very receptive to feedback; even if there were several in a row that I didn't speak to anyone about, they would be like "Are we OK? How are we looking as a table?"

As data table 3 documents, scoring at Table C was overturned frequently; yet no other table leader reported such spontaneous requests for recalibration and discussion. Is it not possible that these raters knew on some level that their scoring was problematic, even if they could not quite have said

how or why they knew? What if this table leader had been trained to listen for such questions as a potential source of information—in this case, as her raters' attempts to signal to her that they felt something was wrong?

Chief readers might also adjust conventional models for adjusting severe or lenient scoring. Such models rely on a unitary and outmoded construct of the rater's mind and miss the table as a microecology of scoring. This modernist interpretation may help explain raters' vulnerability to post-intervention hypercorrection or a retreat to prior scoring behaviors (see Congdon and McQueen). Instead of pulling a rater aside to discuss a pattern of aberrant scores, table leaders might invite a table-level discussion about the causes and possible interpretations of a discrepant score, interpreting that pattern of scoring within a model that emphasizes the communal nature of reading (Knoch "Rating").

Second, while holistic scoring procedures have long emphasized highly orchestrated initial training (see Bejar 4-5), there appear to be few protocols for attending to questions that arise during reading. For example, recall (as discussed above) that Rater A3 had been asked to adjudicate one of Rater D3's nonadjacent scores and that he had no reason to suspect that D3 was scoring so severely. Weighing both scores as potentially legitimate, A3 develops an important insight about the instrumental validity of the scoring sheet.

> A3: This author . . . was in her reflective letter misdirecting readers' attention from the things she was actually demonstrating that she did well. And so, if you were to take her reflection simply at face value, she was very convincing that she didn't do anything well. But reading her other papers, I think it was actually defensible that she was actually doing these things; she just wasn't entirely sure how to reflect upon them effectively.
>
> Dryer: So she was potentially disadvantaged by this particular scoring protocol?
>
> A3: I do think so.

Chief readers may overestimate the durability of the effect of calibration and the transparency of rubrics. Recent insights into the brittleness of conceptual gains made by novice teachers (Reid, Estrem, and Belcheir) point to the powerful longevity of prior dispositions (especially those tacitly operating on the field level). A questionnaire (like the one we put to the PASS Raters) administered before an assessment could help chief readers make better decisions about where to invest the limited resources of calibration;

under ideal conditions, the results of such a questionnaire would be shared and discussed with raters themselves before calibration begins.

Third, assessment designs—including scoring guides—may assume more shared agreement about the meaning of key terms than actually exist at the tables. Even the table leader focus group acknowledged the tenuousness of their grasp on the writing construct:

> TL-D: The assumption was that we were going to agree with all of the scores and the way that they were seen by whoever selected them. . . . We should have at least gotten those samples ahead of time so that we could have read them and had a *discussion* about them amongst ourselves. We had the answers but without the rationale. (Original emphasis)

A weakening alignment with the construct of an assessment has usually been framed as "severity drift," and both human and algorithmic techniques are routinely deployed to spot and correct it (e.g. Hoskens and Wilson). But an ecological construct suggests that weakening alignment could be reframed as an occasion for renegotiating interpretations of the scoring guide. In this context, it is compelling that both focus groups argued for, as rater D2 put it, "structured times when readers need to feel they can get up and go." Rater C1 agreed, laughing, "Where they *make* you get up and go!" At the table leader focus group, TL-C said almost precisely the same thing:

> You have to structure times when readers need to feel they can get up and go—'OK, we're all breaking now.' Because someone would come back and they'd be like 'Oh, where am I, I've got to hurry up with this,' and I thought that might have affected us.

To be sure (and not to put too fine a point on it), "going" probably means toilet breaks. But that can't be all that raters meant, given that they independently spoke of "structured time." A variety of collective experiences (even waiting in line for lunch) might offer chances to renegotiate their experiences of the assessment: opportunities to vent, contextualize pet peeves, to share tacit definitions of a term such as *process*, or to compare impressions like rater B3's, which she probably would not have articulated for herself had she not been interviewed: "I was surprised by how many of the portfolios fell right in the center, and I thought for a while 'did I just get all of the average ones?'"

It appears that assessment administrators should expend as much capital as they can spare on time: time for raters to excavate tacit assumptions, time to explore the premises of the assessment, time to work toward a more closely shared definition of key terms (the exit survey alone suggests a wide range of private definitions of the term *rhetorical knowledge*). Material con-

ditions alone make such time costly (raters must be paid; schedules are difficult to coordinate), and cultural conditions conspire against enacting an ecological construct of writing assessment: social conventions default toward hierarchies of surveillance and control; social imperatives for "objective" scoring discourage negotiation and discussion, even when such discussion might increase the reliability of scoring.

Finally, moving from the dynamics operating within an assessment to the reporting of its scoring, we suggest that the use-validity of assessment data is enhanced by reporting the scoring protocol and its relation to field-level criteria, making a routine practice of disaggregating scores by table and rater, and providing those profiles alongside the scoring profile of the room. It might be objected that this practice would undermine the years of effort spent gaining the confidence of external stakeholders for reliable direct assessments of writing, but stakeholders who never see table-level distributions alongside room-level distributions are less likely to see the full complexity of the construct assessed. Accordingly, they may believe that raters' scores (and thus the decisions made on the basis of those scores) are less consequential than they are; they may also be tempted to believe that writing assessment can be outsourced with little harm to the process. Users of the information derived from large-scale, program-wide assessments should be aware of the degree to which a range of social systems introduce volatility into any large-scale assessment of student writing. Users who are more aware of the full complexities that produce scores are likely to be more circumspect in their use of that information than history suggests has been the case so far.[11]

Notes

1. For much of the twentieth century, writing assessments operated with an assumption of generalized writing ability—i.e., an ability to answer selected-response items about grammatical conventions or to produce an impromptu belletristic essay were sufficient proxies for writing ability. In spite of subsequent research documenting dramatic differences in the conventions of different genres as well as writers' difficulties in transferring writing ability across genres, this assumption still operates in the most powerful US large-scale writing assessments (as their overgeneralized names suggest: the former "SAT Writing" test and the still-current "AP Language and Composition" and the "ACT Writing Test").

2. We have adopted MacMillan's framework of four "concentrically embedded contextual layers" suited to ecological research in writing with terms suited to assessment (335). *Rater, Table*, *Room*, and *Field* correspond to *micro-*, *meso-*, *exo-*, and *macrosystems* respectively.

3. A pseudonym.

4. Another pseudonym.

5. A pseudo-acronym.

6. We were included in team meals to give us access to conversations among the chief reader and table leaders but otherwise paid all our own expenses.

7. This research was done prior to the publication of WPA Outcomes Statement 3.0.

8. Another pseudonym.

9. Sinclair and the Jennings-Baker representative both emphasized the academic nature of the scoring, possibly to obviate concerns about the corporate sponsorship of the assessment affecting the integrity of the second purpose above. "We are," Sinclair said to the raters, "not being asked to do anything to favor Jennings-Baker but rather to come out with as careful a set of data on the material we have as possible and let the conclusions fall where they may."

10. We cannot conclude anything about the degree to which they belonged in the field before the assessment because we were unable to distribute these questionnaires before the assessment began.

11. This study was judged Exempt, Category 2 by the Institutional Review Boards of the University of Maine (application #2009-12-01) and Louisiana State University (IRB# E4870).

Works Cited

Baker, Beverly A. "Playing with the Stakes: A Consideration of an Aspect of the Social Context of a Gatekeeping Writing Assessment." *Assessing Writing* 15.3 (2010): 133–53. Print.

Barkaoui, Khaled. "Do ESL Raters' Evaluation Criteria Change with Experience?" *TESOL Quarterly* 44.1 (2010): 31–57. Print.

Barritt, Loren, Patricia Stock, and Francelia Clark. "Researching Practice: Evaluating Assessment Essays." *College Composition and Communication* 37.3 (1986): 315–27. Print.

Bazerman, Charles. *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Madison: U of Wisconsin P, 1988. Print.

Behizadeh, Nadia, and George Englehard. "Historical View of the Influences of Measurement and Writing Theories on the Practice of Writing Assessment in the United States." *Assessing Writing* 16 (2011): 189–211. Print.

Bejar, Issac I. "Rater Cognition: Implications for Validity." *Educational Measurement: Issues and Practice* 31.3 (2012): 2–9. Print.

Berlin, James. *Rhetoric and Reality: Writing Instruction in American Colleges, 1900-1985*. Carbondale: SIUP, 1987. Print.

Brennan, Robert L., ed. *Educational Measurement*. 4th ed. Westport, CT: American Council on Education and Praeger, 2006. Print.

Broad, Bob. "Pulling your Hair Out: Crises of Standardization in Communal Writing Assessment." *Research in the Teaching of English* 35.2 (2000): 213–60. Print.

Broad, Bob, Linda Adler-Kassner, Barry Alford, Jane Detweiler, Heidi Estrem, Susanmarie Harrington, Maureen McBride, Eric Stalions, and Scott Weeden, eds. *Organic Writing Assessment: Dynamic Criteria Mapping in Action*. Logan: Utah State UP, 2009. Print.

Callahan, Susan. "Responding to the Invisible Student." *Assessing Writing* 7.1 (2000): 57–77. Print.

Carrell, Patricia L. "The Effect of Writers' Personalities and Raters' Personalities on the Holistic Evaluation of Writing." *Assessing Writing* 2.2 (1995): 153–90. Print.

Colombini, Crystal B., and Maureen McBride. "'Storming and Norming': Exploring the Value of Group Development Models in Addressing Conflict in Communal Writing Assessment." *Assessing Writing* 17.4 (2012): 191–207. Print.

Condon, William. "Large-Scale Assessment, Locally-Developed Measures, and Automated Scoring of Essays: Fishing for Red Herrings?" *Assessing Writing* 18.1 (2013): 100–08. Print.

Congdon, Peter J., and Joy McQueen. "The Stability of Rater Severity in Large-Scale Assessment Programs." *Journal of Educational Measurement* 37.2 (2000): 163–78. Print.

Cooper, Marilyn. "Rhetorical Agency as Emergent and Enacted. " *College Composition and Communication* 62.3 (2011): 420-49.

Crisp, Victoria. "An Investigation of Rater Cognition in the Assessment of Projects." *Educational Measurement: Issues and Practices* 31.3 (2012): 10–20. Print.

Diederich, Paul. *The Measurement of Growth in English*. Ubana: NCTE, 1974. Print.

Dryer, Dylan B. "At a Mirror, Darkly: The Imagined Undergraduates of Ten Novice Composition Instructors." *College Composition and Communication* 63.3 (2012): 420–52. Print.

Dryer, Dylan B. "Scaling Writing Ability: A Corpus-Driven Inquiry." *Written Communication* 30.1 (2013): 3–35. Print.

Elliot, Norbert. *On a Scale: A Social History of Writing Assessment in America*. New York: Peter Lang, 2005. Print.

Elliot, Norbert, Vladimir Briller, and Kamal Joshi. "Portfolio Assessment: Quantification and Community." *Journal of Writing Assessment* 3.1 (2007): 5–30. Print.

Hambleton, Ronald K., and Mary J. Pitoniak. "Setting Performance Standards." Brennan 433–70.

Hamp-Lyons, Liz. "Rating Nonnative Writing: The Trouble with Holistic Scoring." *TESOL Quarterly* 29.4 (1995): 759–62. Print.

—. "Writing Assessment: Shifting Issues, New Tools, Enduring Questions." *Assessing Writing* 16.1 (2011): 3–5. Print.

Hamp-Lyons, L., and Sheila P. Mathias. "Examining Expert Judgments of Task Difficulty on Essay Tests." *Journal of Second-Language Writing* 3.1 (1994): 49–68. Print.

Harrington, Susanmarie. "Introduction." *The Outcomes Book: Debate and Consensus after the WPA Outcomes Statement*. Ed. Susanmarie Harrington, Keith Rhodes, Ruth Overman Fischer, and Rita Malenczyk. Logan: Utah State UP, 2005. xv–xix. Print.

Harsch, Claudia, and Guido Martin. "Adapting CEF-descriptors for Rating Purposes: Validation by a Combined Rater Training and Scale Revision Approach." *Assessing Writing* 17.4 (2012): 228–50. Print.

Haswell, Richard, and Janis Tedesco Haswell. "Gender Bias and Critique in Student Writing." *Assessing Writing: A Critical Sourcebook*. Ed. Brian Huot and Peggy O'Neill. Boston: Bedford/St. Martin's, 2009. 387–434. Print.

Heath, Christian, and Paul Luff. *Technology in Action.* Cambridge UP, 2000. Print.

Hoskens, Machteld, and Mark Wilson. "Real-Time Feedback on Rater Drift in Constructed-Response Items: An Example from the Golden State Examination." *Journal of Educational Measurement* 38.2 (2001): 121–45. Print.

Huot, Brian. *(Re)Articulating Writing Assessment for Teaching and Learning*. Logan: Utah State UP, 2002. Print.

Huot, Brian, and Michael Neal. "Writing Assessment: A Techno-History." *Handbook of Writing Research.* Ed. Charles A. MacArthur, Steve Graham, and Jill Fitzgerald. New York: Gilford, 2006. 417-32. Print.

Huot, Brian, Peggy O'Neill, and Cindy Moore. "A Usable Past for Writing Assessment." *College English* 72.5 (2010): 495–517. Print.

Kane, Michael. "Validation." Brennan 17–64.

Knoch, Ute. "Investigating the Effectiveness of Individualized Feedback to Rating Behavior—A Longitudinal Study." *Language Testing* 28.2 (2011): 179–200.

—. "Rating Scales for Diagnostic Assessment of Writing: Where Should the Criteria Come From?" *Assessing Writing* 16.2 (2011): 81–96. Print.

Lane, Suzanne, and Clement A. Stone. "Performance Assessments." Brennan 387–432.

Lumley, Tom, and TF McNamara. "Rater Characteristics and Rater Bias: Implications for Training." *Language Testing* 12.1 (1995): 238–57. Print.

Maid, Barry, and Barbara D'Angelo. "Is Rhetorical Knowledge the Über-Outcome?" *The WPA Outcomes Statement: A Decade Later*. Ed. Nicholas Behm, Gregory R. Glau, Deborah H. Holdstein, Duane Roen, and Edward M. White. Anderson: Parlor P, 2013. 257–270. Print.

MacMillan, Stuart. "The Promise of Ecological Inquiry in Writing Research." *Technical Communication Quarterly* 21.4 (2012): 346–61. Print.

McCutchen, Deborah, Paul Teske, and Catherine Bankston. "Writing and Cognition: Implications of the Cognitive Architecture for Learning to Write and Writing to Learn." *Handbook of Research on Writing: History, Society, School, Individual, Text*. Ed. Charles Bazerman. New York: Lawrence Erlbaum Assoc, 2008. 451–70. Print.

Moffett, James. *The Universal Schoolhouse*. San Francisco: Jossey-Bass, 1994. Print.

Myford, Carol, and Edward W. Wolfe. "Monitoring Rater Performance over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use." *Journal of Educational Measurement* 46.4 (2009): 371–89. Print.

Peckham, Irvin. *Going North, Thinking West: The Intersections of Social Class, Critical Thinking, and Politicized Writing Instruction*. Logan: Utah State UP, 2010.

—. "Online Challenge Versus Offline ACT." *College Composition and Communication* 61.4 (2010): 718–45. Print.

Poe, Mya, and Asao Inoue. "Introduction." *Race and Writing Assessment.* Ed. Mya Poe and Asao Inoue. New York: Peter Lang, 2012. 1–19. Print.

Powers, Donald E., Mary E. Fowles, Marisa Farnum, and Paul Ramsey. "'Will They Think Less of My Handwritten Essay If Others Word Process Theirs?' Effects on Essay Scores of Intermingling Handwritten and Word-Processed Essays." *Journal of Educational Measurement* 31.3 (1994): 220–33. Print.

Prior, Paul. "Tracing Processes: How Texts Come into Being." *What Writing Does and How It Does It*. Ed. Charles Bazerman and Paul Prior. New York: Routledge, 2003. 167–200. Print.

Reid, E. Shelley, and Heidi Estrem, with Marcia Belcheir. "The Effects of Writing Pedagogy Education on Graduate Teaching Assistants' Approaches to Teaching Composition." *WPA: Writing Program Administration* 36.1 (2012): 30-73. Print.

Singer, Nancy R., and Paul LeMahieu. "The Effect of Scoring Order on the Independence of Holistic and Analytic Scores." *Journal of Writing Assessment* 4.1 (2011); n. pag. Web. 23 October 2011.

Suto, Irenka. "A Critical Review of Some Qualitative Research Methods Used to Explore Rater Cognition." *Educational Measurement: Issues and Practice* 31.3 (2012): 21–30. Print.

Vaughan, Caroline. "Holistic Assessment: What Goes on in the Raters' Minds?" *Assessing Second Language Writing in Academic Contexts*. Ed. Liz Hamp-Lyons. Norwood: Ablex, 1991. 111–25. Print.

Wardle, Elizabeth, and Kevin Roozen. "Addressing the Complexity of Writing Development: Toward an Ecological Model of Assessment." *Assessing Writing* 17.2 (2012): 106–19. Print.

Weigle, Sara C. "Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative and Qualitative Approaches." *Assessing Writing* 6.2 (1999): 145–78. Print.

White, Edward M. "The Scoring of Writing Portfolios: Phase 2." *College Composition and Communication* 56.4 (2005): 581–600. Print.

Wiseman, Cynthia S. "Rater Effects: Ego Engagement in Rater Decision-Making." *Assessing Writing* 17.3 (2012): 150–73. Print.

Wolfe, Edward W. "The Relationship between Essay Reading Style and Scoring Proficiency in a Psychometric Scoring System." *Assessing Writing* 4.1 (1997): 83–106. Print.

—. "Uncovering Rater's Cognitive Processing and Focus Using Think-Aloud Protocols." *Journal of Writing Assessment* 2.1 (2005): 37–56. Print.

Wolfe, Edward W., Cei-Wen Kao, and Michael Ranney. "Cognitive Differences in Proficient and Nonproficient Essay Scorers." *Written Communication* 15.4 (1998): 465–92. Print.

Wolfe, Edward W., and Aaron McVay. "Application of Latent Trait Models to Identifying Substantively Interesting Raters." *Educational Measurement* 31.3 (2012): 31–37. Print.

Zhang, Mo. "Contrasting Automated and Human Scoring of Essays." ETS, 2013. Print.

Dylan B. Dryer is Associate Professor of Composition Studies at the University of Maine. Since graduating from the University of Wisconsin–Milwaukee, he has been exploring the capacities for and consequences of genre uptake, a topic with implications for writing assessment, teacher training, identity formation, and the persistence of social institutions generally. His research articles, one of which won the 2013 CCCC Braddock Award, feature usability studies, mixed-method qualitative investigations, and corpus analysis. He is currently guest editing a special issue of *Composition Forum* on the past, present, and possible futures of rhetorical genre studies.

Irvin Peckham is writing program administrator at Drexel University. He was director of the writing program for ten years at Louisiana State University and for four years at the University of Nebraska, Omaha. His research interests are personal writing, writing assessment, and the intersections of social class and writing instruction. He is the author of *Going North Thinking West: The Intersections of Social Class, Critical Thinking, and Politicized Writing Instruction*, co-author with Edward White and Norbert Elliot of *Very Like a Whale: The Evaluation of Writing Instruction* (in press), and articles in several edited collections and in *WPA: Writing Program Administration, Composition Studies, Pedagogy, Computers and Composition, English Journal*, and *College Composition and Communication*.