Researching Teacher Evaluation of Second Language Writing via Prototype Theory

Richard H. Haswell
Haas Professor Emeritus
Texas A&M University, Corpus Christi, TX, USA
rhaswell@grandecom.net

[Author's note, 15 January 2007. This piece was presented as an invited address at the third Second-Language Writing Conference, Purdue University, 12 October 2002. It was published in Paul Kei Matsuda and Tony Silva (Eds.), *Second Language Writing Research: Perspectives on the Process of Knowledge Construction* (Erlbaum, 2005), pp. 105-120. The version here adds Table 8.5, inadvertently omitted from the published piece, and corrects several references to other tables in the text. Internal reference to chapters allude to the Matsuda and Silva volume.]

I begin with a puzzle, a famous puzzle, one that I will solve in two ways. A pair of bicyclists, A and B, are 20 miles apart. They are moving toward each other in a straight line, each at 10 miles an hour. A fly takes off from the front rim of bicycle A and heads in a straight line toward bicycle B, flying at 15 miles an hour. As soon as it touches the front rim of bicycle B, it heads back to bicycle A, then back to bicycle B, and so on. (It may help to picture the fly as an industrious dean.) When the two bicycles meet and presumably the fly is squashed between the front tires, how far has it flown?

As I say, there are at least two ways to arrive at the answer, which happens to be 15 miles. One is to measure each leg of the fly's journey, taking into account the diminishing distance between bicycles with each leg, and to sum the infinite series—a method assisted greatly by calculus. The other way can be accomplished by people such as myself who never mastered calculus. It's an armchair method that reasons as follows. If the two bicyclists are 20 miles apart, each will have traveled 10 miles when they meet. If they are traveling at 10 miles an hour, they will meet in 1 hour. If the fly is flying at 15 miles an hour, then in 1 hour it will have traveled 15 miles.

I will return to this puzzle at the end of the chapter with an estimate of how far I have traveled. In the meantime, it gives me a place to start. Moving to the second way of solving it may entail what cognitive scientists used to call a frame shift, or what creativity theorists used to call a re-orientation. A mental blink switches us from miles covered to hours spent. I want to argue that today research into evaluation of student writing in general and teacher research into evaluation in particular could use a blink. Researchers are frame stuck and need a re-orientation to change the way they solve problems. I am speaking of both evaluation of first language writing and evaluation of second language writing. Both fields are stuck in the same frame—not a surprise because both cycled through the last century so much in tandem, at least in terms of research methodology.

In this chapter, I am going to look directly at that uncomfortable and sometimes seamy side of L2 and L1 composition called proficiency testing. One instance of my topic is Linda Blanton's student, Tran, and the examination hurdle he had to jump to get his degree (chap. 11, this volume). I am further focusing in on one practice of such testing—the construction and use of evaluative categories—categories such as "holistic level 2," "proficient," "band 3," and "ready

for advanced composition." Actually, I will focus twice: once to critique these kinds of categories and again to recommend a quantified method by which to research them. Perhaps my procedure fits Tony Silva's definition of critical rationalism (chap. 1, this volume). Certainly my critique extends beyond standardized testing and includes teachers, researchers, and students because all of them, every day, use value-laden categories in connection with student writing.

Dwight Atkinson asks researchers to try to know people "on their own terms" (chap. 4, this volume). But do people themselves know their terms, and do researchers know the terms by which they try to know people? These "terms" are almost always categories. The bind is that categories are normally obscure to the user of the categories, certainly the inner workings of them. This chapter is simply recommending to researchers "a tool for seeing the invisible," more exactly for seeing "the role of invisibility in the work that categorization does in ordering human interaction" (Bowker & Star, 2000, p. 5). The tool it recommends is not the only way out of the frame-stuck position of L1/L2 evaluation research, but it should be better known. As a tool of current critical rationalism, it is familiar to nearly every social science field except for composition studies. In gist the research method will help shift attention from the application of writing criteria to the grounds for writing criteria—a shift enabled with categorization theory, and a shift requiring a turn from the standardized to the local. It is a simple mental click, but every part of it turns out to be very complex. This chapter covers two somewhat separate steps of the shift. Part I looks at the way standardized testing categorizes essay writing criteria, and Part II looks at the way faculty could differently, and perhaps better, categorize essay writing main traits.

Part I: How big testing operations categorize essay writing skills

Historically, commercial testing firms have had a major hand in constructing the evaluative frames current among teachers and researchers. Consider the scoring sheet from Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey's (1981) *Testing ESL Composition: A Practical Approach* (Fig. 8.1). The authors call it the "ESL Composition Profile." Since this scoring guide was published in 1981, it has proved very popular. It, or its offspring, will be familiar from workshop handouts or Xeroxes left behind in faculty coffee rooms. In its main features, it is no different than dozens of similar guides by which raters have decided, and continue to decide, the academic fate of thousands upon thousands of second language students. These main features are:

1. A limited number of basic criteria or main traits (e.g., content, organization, vocabulary, language use, and mechanics).

2. A fitting of each trait into a proficiency scale, the levels of which are also small in number and usually homologous or corresponding (e.g., 1, 2, 3, or 4 for each trait).

3. A breakdown of each trait into subtraits, which are also small in number and homologous or corresponding. See Table 8.1, which teases out the subtraits of the main trait *content* in Jacobs et al. There are four subtraits

ESL COMPOSITION PROFILE

STUDENT DATE TOPIC

	SCORE	LEVEL	CRITERIA	COMMENTS
CONTENT		30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic	
		26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail	
		21-17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic	
		16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate	
ORGANIZATION		20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/supported • succinct • well-organized • logical sequencing • cohesive	
		17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing	
		13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development	
		9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate	
VOCABULARY		20-18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register	
		17-14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured*	
		13-10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured*	
		9-7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate	
LANGUAGE USE		25-22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions	
		21-18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured*	
		17-11	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • *meaning confused or obscured*	
		10-5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate	
MECHANICS		5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing	
		4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured*	
		3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • *meaning confused or obscured*	
		2	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate	

TOTAL SCORE READER COMMENTS

Figure 8.1: ESL Composition Profile (Jacobs et al., 1981, p. 30), reprinted with permission of the publisher

> each with corresponding levels: knowledge of the topic, substance, development of the topic, and relevance. The homology, it should be noted, does not allow for a writer who has "a limited knowledge" of the topic, yet applies what little she or he knows in a way that is "relevant" to the topic

This hidden feature of homology, very significant, has been little discussed by composition researchers. The "ESL Composition Profile" is lauded because it is just that—a *profile* of the student, not a categorization of the student. It encourages an evaluation of student proficiency that is complex, perhaps recording high accomplishment in content, but low in mechanics—a complexity that befits writers who often show uneven writing skills in a second language. In this the profile seems to contrast with holistic scoring methods, which erase this possible unevenness of writing accomplishments in reporting a single score. But in fact the kind of rating that underlies the "ESL Cornposition Profile" is identical to holistic rating. The "Profile" just asks the rater to perform the holistic five times. In short—this is the emphasis I put on it—both methods of scoring ask the rater to apply the same *kind of categorization*.

Table 8.1

Breakdown of Subtrait Levels for the Main Trait of *Content* in Figure 8.1 (Jackobs et al., 1981, p. 30), Showing Homology of Levels

Level	*Subtrait 1*	*Subtrait 2*	*Subtrait 3*	*Subtrait 4*
Points	*Knowledge*	*Substance*	*Development*	*Relevance*
30-27	Knowledgeable	Substantive	Thorough development of thesis	Relevant to assigned topic
26-22	Some knowledge of subject	Adequate range	Limited development of thesis	Mostly relevant to topic, but lacks detail
21-17	Limited knolwedge of subject	Little substance	Inadequate development [of topic]	Inadequate [relevance to] topic
16-13	Does not show knolwedge of the subject	Nonsubstantive	Not enough to evaluate	Not pertinent

I will return to this fact later in this chapter, but first it is worth observing how the features of holistic or profile scoring lend themselves to the kind of evaluation research that has dominated L2 and L1 composition studies for decades. The limited number of traits allows comparison of group rating behavior, perhaps contrasting the way native and non-native faculty evaluate ESL essays. The scaling of traits and subtraits allows study of rater reliability along with the development of training methods that produce high interrater reliability coefficients needed to defend commercial testing or research studies. The reduction of uneven and otherwise complex

writing proficiency to units, and the internal ordering of traits or subtraits as homologous and mutually exclusive, allow the generation of empirical outcomes useful in research, placement, and program validation.

Here comes the blink. What happens when the "ESL Composition Profile" sheet is re-oriented with a new question? The five main traits, whose names in fact are oriented differently than the rest of the words on the sheet (see Fig. 8.1)—where did they come from? The question does not ask how well they function as parts of a performance-evaluation mechanism. I am asking who put these main traits into the mechanism and why, not how well they work once put there. Factor analysis of scores produced by this mechanism, for instance, might eliminate one of these traits if it contributes no unique information to the profile, but it could not find another and better trait for replacement. The question asks why content, organization, vocabulary, language use, and mechanics and not creativity, logic, suspense, tradition, shock-appeal, humor, cleverness—and the second list could go on.

The question is not trivial or irrelevant. The criteria not chosen shape the outcomes as much as those that are chosen. Now it happens that the origin of these five main traits of the "ESL Composition Profile" is known, and the providence may be surprising. They came from grades and marginal comments written on student homework. The graders and commenters included a few teachers, but most were social scientists, natural scientists, workplace editors, lawyers, and business executives. None of them had TESOL experience. The writers were first-year students at Cornell, Middlebury College, and the University of Pennsylvania, probably none of them second language students. This was in 1958 (Diederich, 1974; Diederich et al., 1961). Three researchers at Educational Testing Service (ETS) factored the commentary, passed the factoring on to a colleague of theirs at ETS, Paul Angelis, who passed it on to the authors of thc "ESL Composition Profile" (Jacobs et al., 1981). In the relay, one of the original five factors, flavor, got dropped, and another, wording, got divided into vocabulary and language use, but no new factors were added. So the main criteria of a popular L2 essay rating method were derived not from L2 essays, nor from L2 teachers, nor much from teachers at all.

Main traits of other long-lived second language writing tests probably have equally troubling and mysterious histories (Table 8.2). I do not know the archaeology of these categorizations. To find out would make a fascinating study, but I suspect many of them originated with a certain amount of blithe. Certainly the following rationale by David P. Harris is blithely expressed. Harris was the project director of the TOEFL exam from 1963 to 1965, and his comment appears in his 1969 book, *Testing English as a Second Language*: "Although the writing process has been analyzed in many different ways, most teachers would probably agree in recognizing at least the following five general components: Content, Form, Grammar, Style, Mechanics" (pp. 68-69).

Part II: How researchers and teachers might categorize essay writing skills

In fact most teachers do not agree, certainly not ESL teachers, and certainly not on the assumption, which Harris implies, that these five components are equally important. The question is how can teacher/researchers find out what teachers do agree on? I am turning to teacher/researchers because I do not have much faith that the giant testing firms will ever change their ways. There are many inquiry methods, "tools for seeing the invisible," to ferret out main

TABLE 8.2

Main Traits of Scoring Rubrics for Six Tests of ESL Writing

Test	*Trait* *Organization*
Test in English for Educational Purposes (Associated Examining Board)	*Content* *Organization* *Cohesion* *Vocabulary* *Punctuation* *Spelling*
Certificate in Communicative Skills in English (Royal Society of Arts/University of Cambridge Local Examinations Syndicate)	*Accuracy* [of mechanics] *Appropriacy* *Range* [of expression] *Complexity* [organization and cohesion]
Test of Written English (Educational Testing Service)	*Length* *Organization* *Style* *Grammar* *Sentences*
Michigan English Language Battery	*Topic development* *Sentences* *Organization/coherence* *Mechanics*
Canadian Test of English for Scholars and Trainees	*Content* *Organization* *Language use*
International English Lnguage Testing System	*Register* *Rhetorical organization* *Style* *Content*

traits, ranging from rater think-aloud protocols to participant/observer ethnography of rating groups. Some of this inquiry is and has been going on, some of it both massive and complex—calculus solutions. The International Association for the Evaluation of Educational Achievement project of the early 1980s explored 28 criteria (Purves & Takala, 1982), and Grabe and Kaplan's (1996) taxonomy of language knowledge erects 20 main categories encompassing 41 distinct traits, to mention one old and one recent study. The remainder of this chapter offers a less onerous method, an armchair procedure, if you will. It is a procedure based on prototype categorization theory.

To start, let me review the sociocognitive prototype model of categorization. Prototype categorization stands in opposition to classical notions of category definition that centuries of

Aristotelian logic have formed and that seem almost too obvious to question. A classical category of writing evaluation—say "writing mechanics"—has a fixed set of defining features that provide the category with absolute boundaries. Things—say a failure to spell conventionally the word *America*—fall into the category or not. Members within a category, then, belong there equally. An incorrect full stop is no less good an instance of writing mechanics than a misspelling of *America*. Compared to this *classical* categorization, *prototypical* categorization operates quite differently. It does not fix a set of defining features, but rather organizes itself around a best example or prototype. Members of the category stand closer or further from this prototype. Members or parts of a category are not equal—they have different degrees of centrality or "goodness of part" (this quality of prototype categories is sometimes called *graded structure*). Some members are so marginal that they may be closer to the prototype of another category. Hence, prototypical categories do not have boundaries. They overlap with other categories, and instances may belong equally to two different categories. The title of Kafka's novel, which spells *America* with a "k," may belong equally to mechanics, style, or creativity. Other kinds of categorization have been explored, of course, and argument continues over their use and interrelationships in the evaluation of human performance (for reviews of prototype theory, see Hampton, 1993; Haswell, 1998; Lakoff, 1987).

But there should be little argument that commercial essay testing must treat prototypical categorization as anathema. Consider again the "ESL Composition Profile" (Fig. 8.1). At no less than three points, the procedure is structured classically: the overall categorization of "writing proficiency" with its five defining features, content, organization, vocabulary, language use, and mechanics; the "mastery levels," again with absolute boundaries, as the scale points make clear; and the subtraits, separate but homologous (Table 8.1).

Prototypical categorization would destroy this system of evaluation. It would certainly scrap the overall configuration of the "ESL Composition Profile" as all compartments and right angles, a fearful symmetry that, among other things, expresses a fear of overdetermination and overlap. To note just one fear, conceptual slippage would lead directly to a slip in the interrater reliability coeffficient.

What's wrong with the Profile's classical way of categorizing as a means of evaluation? Two generations of categorization researchers, who have explored the way prototype categorizing affects evaluation of human performance in nearly every area imaginable, offer a clear and even unforgiving answer. The trouble with classical categorization is that it doesn't account for the way people normally categorize. People do judge some misspellings of *America* as worse than others, some even as clever and a matter of style, not mechanics, some even as thoughtful and a matter of content and not a matter of mechanics at all. "Every category observed so far," summarizes Lawrence Barsalou (1987), "has been found to have graded structure" (p. 111).

Some researchers in second language acquisition have participated in this robust inquiry into typicality effects. Lindstromberg (1996) works with the teaching of prepositions as graded categories, and Taylor (1995) masterfully treats grammatical and lexical concepts as prototypical. But as far as I know, no one has studied how the human rating or grading of L2 essays categorizes in prototypical ways. Yet so much of what a writing teacher does is categorizing. To put a "B" on a paper is to pigeonhole it in the category of "B work." To read a

placement essay and assign the writer to the second level of an ESL curricular sequence is to categorize the writing as "intermediary work." To judge an ESL writer's exit portfolio from first-year composition as successful is to place it in the basket labeled "ready for advanced composition." These are all acts of categorization, and the last 30 years of psychological and sociological research into the way people categorize would argue that only by a miracle would these acts follow Aristotelian rules, would not show prototype effects.

Let me take the last example—end-of-first-year proficiency—and use it to show what one piece of writing evaluation research based on prototype theory might look like. For this chapter, I ran a preliminary, online study to demonstrate the method. My research question was this: In making decisions about end-of-first-year writing proficiency, do L2 teachers categorize writing traits differently in their evaluation of L2 writers than do L1 teachers L1 writing? For traits, I selected 10 from the Council of Writing Program Administrator's (2000) recent Writing Outcomes document, the 10 that can be most readily inferred from written text (Table 8.3). I applied a

TABLE 8.3

Ten Essay Writing Traits Selected From the Writing Program Administrators Outcomes Statement for First-Year Composition

Short Form	*Description*
Audience	Responds to the needs of the readers
Documentation	Uses appropriate means of documenting the writing
Inquiry	Shows the use of inquiry, learning, thinking
Integration	Integrates their ideas with the ideas of others
Purpose	Focuses on and conveys the purpose for the writing
Situation	Responds appropriately to the rhetorical situation
Sources	Uses sources conventionally and well
Structure	Arrangement or format fits the rhetorical situation
Surface	Is in control of surface features (spelling, etc.)
Voice	Adopts an appropriate voice, tone, formality

matched guise design, with two groups of teachers evaluating the same essay under difference preconceptions. One group, L1 teachers, believed it was written by an L1 writer; the other group, L2 teachers, believed it was written by an L2 writer. Finally, and most crucially, I had each participant first rate each trait (on a 7-point Likert scale) in terms of its prototypicality or "centrality" in their evaluation of first-year writing accomplishment. Note that this teacher-rater generalized judgment of the 10 traits preceded their specific rating of an essay in terms of the traits. In this way, although the procedure ended up generating a profile of one student's essay along a set number of writing traits, just as in the "ESL Composition Profile," it produced further information—prototypical information, namely, how central the raters thought those traits were.

This experiment did not escape some of the problems of online research. Control of participant selection was weak, and the conditions under which participants performed the evaluation could not be regularized. Also the number of participants (43 L2 teachers and 57 L1 teachers) was too low to support thorough inference testing of 10 traits. Therefore, although the experiment was conducted with all the rigor the conditions allowed, I offer the findings only as illustrative. They

do argue the feasibility of prototypical inquiry in L1/L2 research and suggest the method's potential in challenging traditional evaluation research, little of which is based on any other assumptions than those of classical categorization.

Let me first look at the second step of the evaluation, in what seems like rather astonishing support for traditional evaluation (Table 8.4). At this point in the experiment, all teachers had

TABLE 8.4

Rating of One Essay Under Two Preconceptions ("ESL" and "Native"*) by Two Sets of Readers ("ESL," N = 43, L2 Teachers; "Native," N = 59, L1 Teachers) on a Likert Scale (7 = *high proficiency*, 1 = *low proficiency*)

ESL		*Native*	
Scale Mean *(SD)*	Trait	Scale Mean *(SD)*	Trait
5.54 *(1.05)*	Voice	4.44 *(1.61)*	Documentation
5.35 *(1.43)*	Documentation	4.41 *(1.46)*	Voice
4.65 *(1.73)*	Purpose	3.73 *(1.74)*	Purpose
4.61 *(1.53)*	Inquiry	3.66 *(1.40*	Inquiry
4.30 *(1.34)*	Audience	3.60 *(1.26)*	Audience
3.98 *(1.63)*	Sources	3.53 *(1.48)*	Structure
3.95 *(1.45)*	Structure	3.53 *(1.71)*	Sources
3.88 *(1.62)*	Situation	3.34 *(1.21)*	Situation
3.88 *(1.31)*	Surface	3.27 *(1.57)*	Integration
3.84 *(1.63)*	Integration	3.05 *(1.56)*	Surface

*ESL readers assumed the writer was NNS, from South Korea; native readers assumed the writer was NS, from the U.S. midwest.

been shown the same student essay and asked to rate it in terms of the 10 WPA Outcomes traits. As I have said, the 43 teachers with L2 teaching experience had been led to believe that the essay was written by an ESL student. They understood that the writer was born in Korea and immigrated to the United States 3 years ago. The other 57 teachers, with little or no L2 teaching experience, were led to believe that the essay was written by an NES student, born and raised in the midwest of the United States. Yet the two groups generated evaluation profiles—rankings of the accomplishment of the essay along the 10 WPA traits—that are remarkably similar. On 10 traits and at any point, the group means do not differ by more than one rank. L1 and L2 writing teachers, as rating groups, seem to concur on this essay's writing success regardless of their differing presuppositions about the writer's language status.

Another finding shown here (also shown in Table 8.4) is even more consistent, although it has been reported before in ESL research (see Silva, 1989, for a review). With every WPA trait, the ESL-assumed writer is rated more highly than the native-language-assumed writer. In the online discussion among the participants that followed my presentation of this finding, several L1 teachers suggested that the finding shows the generosity of L2 teachers in rating L2 writing. But it is more reasonably explained in terms of the possible parameters of accomplishment imagined by the two rater groups. The L1 teachers were locating the essay within the typical range of end-of-first-year writing performance of L1 writers and the L2 teachers within the range of L2

writers. In effect the two groups were applying the Likert scale (least proficient at one end, most proficient at the other) to two different imagined populations of student writers. But however the difference in leniency is explained, Table 8.4 seems to show L1 and L2 teachers agreeing on the 10 WPA traits in terms of essay accomplishment with remarkable consistency.

But when we look at the way the two teacher groups judged these 10 traits in terms of centrality or prototypical goodness of part, we see that maybe this L2-L1 teacher concordance is deceptive and hides some deep disagreements. The disagreements are over the internal categorization of these traits vis-à-vis language status. As I mentioned earlier, the two groups had rated the same 10 traits according to their centrality as measures of end-of-first-year proficiency. One group rated them for L2 writers, the other for L1 writers. Table 8.5 shows the group means for

Table 8.5

Judged Centrality of Ten Essay-Writing Traits in Terms of End of First Year at College for Two Assumed Groups of Writers, ESL (N = 43) and Native Speakers (N = 59)* (*Likert scale*: 7 = most central; 1 = most marginal)

ESL		*Native*	
Scale Mean *(SD)*	Trait	Scale Mean *(SD)*	Trait
6.49 *(0.89)*	Purpose	6.61 *(0.97)*	Purpose
5.95 *(1.07)*	Situation	5.98 *(1.15)*	Situation
5.86 *(1.00)*	Structure	5.92 *(1.19)*	Structure
5.62 *(1.23)*	Integration	5.80 *(1.31)*	Documentation
5.54 *(1.18)*	Audience	5.80 *(1.03)*	Surface
5.49 *(1.49)*	Inquiry	5.73 *(1.35)*	Sources
5.30 *(1.51)*	Documentation	5.66 *(1.35)*	Audience
5.26 *(1.65)*	Sources	5.63 *(0.89))*	Voice
5.26 *(1.26)*	Voice	5.63 *(1.23)*	Inquiry
4.74 *(1.65)*	Surface	5.63 *(1.43)*	Integration

•With *ESL* students, judges assumed that they were L2 writers. With *Native* students, judges assumed that they were L1 writers

each trait on a 7-point Likert scale, with 7 being the *most central*, 1 the *least central*. The main prompt for participants is a standard one in prototype research (adapted from Tversky & Hemenway, 1984): "How central is the trait in terms of judging first-year writing accomplishment of the ESL writers [or NES writers] at your university?" After an illustration of "centrality," the prompt ended: "Keep in mind that you are judging the goodness of each trait in showing an ESL [or NES] writer's readiness to exit first-year writing instruction." Comparison of the group means on these 10 traits finds the two groups agreeing on the three most central traits. Whether the student is writing in a first or second language, the prototype or best example of first-year writing proficiency is a well-organized essay with a definite purpose appropriate to the rhetorical situation. At this point, the two groups appear to diverge. Three traits that might fall under Jacobs et al.'s (1981) rubric of mechanics—documentation, surface conventions such as spelling, and appropriate use of sources—are taken as next most central with L1 writers, but

as most marginal with L2 writers. Two traits that reflect depth of thinking—integration of the ideas of others and demonstration of inquiry and learning—are more central in the L2 writers and more marginal in the L1.

But rank alone does not show what I think is the most striking finding here. Rather, it is comparative centrality or goodness of part. Notice in Table 8.5 how each trait is judged more central to first-year writing proficiency for L1 writers and less central for L2 writers. In fact for L2 writers, L2 faculty judged 7 of the 10 traits as less central than L1 faculty judged all 10 of the traits for L1 writers. Figure 8.2, which maps the difference in a way sympathetic to the concept

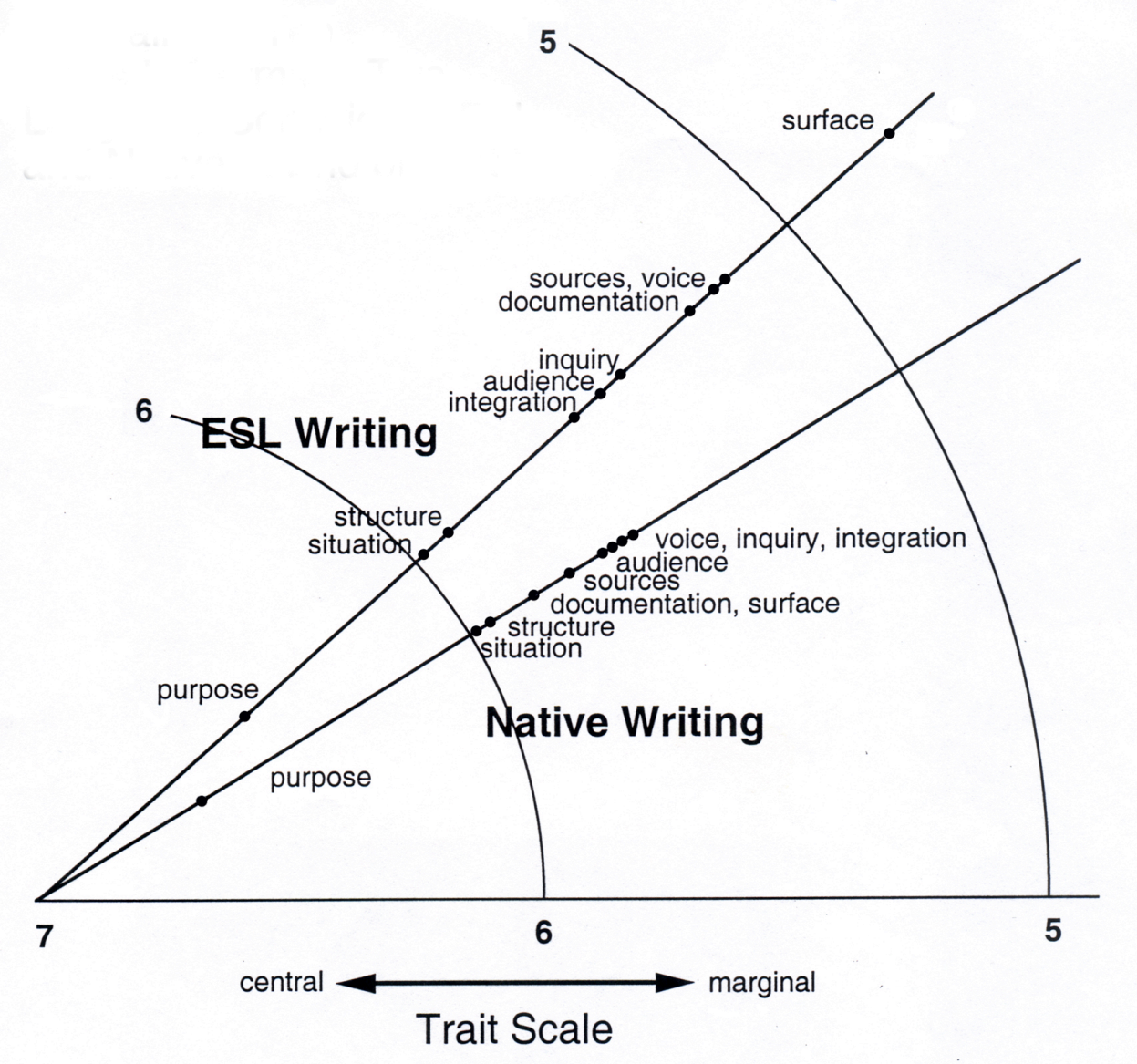

Figure 8.2: Judged Centrality of Ten Traits in Terms of Two Language Conditions: ESL and NES at the End of the Academic Year

of prototypicality, raises disturbing questions. Here is the most dramatic way of putting the findings: On every trait used to judge a particular essay, the presumed L2 writer was rated more highly than the presumed L1 writer; yet on every trait used to judge student writers as a group, the L2 proficiency was judged more marginal than the L1 proficiency. What prototypical categorization giveth, prototypical classification taketh away.

Or maybe it seems that way only at first. Prototypical inquiry returns—as it always will—to my re-orienting question. Where do these main traits come from? How does one interpret this preliminary finding that traits selected by the WPA Outcomes group, all of whom were L1 teachers, do not well fit L2 teachers' conception of L2 proficiency? Do these selected traits, sanctioned by a major organization of U.S. writing program administrators, marginalize second language students? Or does proficiency at a certain point in time, in this case at the end of the first year of college, occupy a less central place in the minds of L2 teachers when they think of L2 students and their writing growth? Or does this set of traits fit the second language teachers' notion of end-of-first-year accomplishment in writing, but just not fit as well? In that case, what more central traits are missing? Table 8.6 provides an initial answer to this final query. I asked research participants—both L1 and L2 teachers—to provide traits that they felt were important in judging first-year proficiency, but were missing from the WPA 10. In several areas, L2 teachers identified proficiencies that contrast with those mentioned by L1 teachers: depth of ideas as opposed to depth of affect or sophistication of argument, interpretation of the issues as opposed to originality of essay construction. The contrasts hint at directions that prototype research models might take the study of ESL evaluation.

At this point, let me insert a comment on inference testing of the centrality data (Table 8.5, Fig. 8.2). Problems in the collection of those data preclude the validation of this particular study with such testing, but I performed some of it anyway, again as a demonstration of method. As an omnibus testing of group performance, a simple *t* test can be used for significant difference between the grand mean of the Likert scale ratings for all 10 traits. Because for each of these traits the mean rating of my L2 raters was more marginal than that of the L1 raters, it is not surprising that the grand mean difference between the two groups was statistically significant (L2 raters, *Mean* 5.55, *SD* 0.78; L1 raters, *Mean* 5.84, *SD* 0.65; *T* 2.02; $p < .046$). *t* tests can also be applied separately to the Likert scale ratings on each of the 10 traits. With this study, two showed significant differences: documentation (L2 raters, *Mean* 5.30, *SD* 1.36; L1 raters, *Mean* 5.78, *SD* 1.19; *T* l.96; $p < .053$), and surface (L2 raters, *Mean* 4.74, *SD* 1.65; L1 raters, *Mean* 5.80, SD 1.03; *T* 3.96; $p < .000$). Finally, intraclass correlations can be run among raters of each group along the 10 trait ratings to judge the degree to which faculty agree in their prototypical structuring of the category "end-of-first-year proficiency." Here concordance among raters in both groups (L2, median *r* .26; L1, median *r* .32) fell within but toward the bottom of the range that has been found in other goodness-of-part studies (cf. Tversky & Hemenway, 1984).

I have just enough space to turn from this exploratory study back to the larger issue—the timeliness of prototype research in L2 writing evaluation. Most important, prototype inquiry can help explore unacknowledged presuppositions not only of teachers, but of students and even commercial test designers. Nor is its potential confined to questions of evaluation. Everywhere there are terms, and the terms are categorizations that could use some deconstruction. To

TABLE 8.6

Traits Judged by College Writing Teachers (L1 N = 43; L2 N = 59) as Useful to the Evaluation of First-Year Writing Proficiency and Missing From the 10 Research Traits

Suggested Traits Common to Both L1 and L2 Teachers	
Syntax	Variation and complexity
Development	Of a thesis statement
Support	With detail and specifics
Cohesion	Transitions, etc.
Revision	Evidence of
Vocabulary	Sophistication and variety
Suggested Traits Unique to L1 Teachers	
Fluency	Putting ideas into words
Ideas	Complexity, elaboration, sophistication, meaningfulness, larger relevance
Interpretation	Of the assignment, or thoughtful response to the research question
Suggested Traits Unique to L2 Teachers	
Fluency	Sustained length
Affectivity	Enthusiasm, curiosity, engagement with the topic, positive attitude toward writing
Argumentation	Multiple perspectives, covering opposing arguments
Originality	Creativity, departures from standard essay structure

mention just one research question raised in this volume, methods of exploring structures of centrality could well provide the grounding that Rosa Manchón argues ethnographers need for the codes through which they analyze participant transcripts (chap. 14, this volume). As just noted, prototype inquiry does not require high-end statistics, either descriptive or inferential. Although it demands the same rigor as any other research, it is a method that can locate new and striking results and do so with armchair calculations.

This brings me back to the puzzle of the bicycles and the fly. As I noted, it can be solved with a complex equation entailing the summing of an infinite series, or it can be solved with a simple division of distance by time. A student who knew both solutions once challenged John von Neumann with the puzzle. Mathematically von Neumann, it may be remembered, had one of the swiftest calculating minds of the last century. He gave the correct answer within a couple of seconds. “Oh, you knew the trick,” said the disappointed student. “What trick?” said von Neumann, “I just summed the infinite series.” There will always be people who find the difficult way easy. Prototype analysis is for them, too.

REFERENCES

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101-140). Cambridge, England: Cambridge University Press.

Bowker, G., & Star, S. (2000). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

The Council of Writing Program Administrators. (200()). *Outcomes statement for first-year composition*. http://www.english.ilstu.edu/Hesse/outcomes.html. Accessed Oct. 2002.

Diederich, P. B. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (ETS Research Bulletin RB-61-15). Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service, ED 002 172).

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. London: Longman.

Hampton, J. A. (1993). Prototype models of concept representation. In I. van Mechelen, J. A. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 67-95). London: Academic Press.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Haswell, R. H. (1998). Rubrics, prototypes, and exemplars: Categorization theory and systems of writing placement. *Assessing Writing*, *5*, 231-268.

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

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Lindstromberg, S. (1996). Prepositions: Meaning and method. *ELT Journal*, *50*, 225-236.

Purves, A. C., & Takala, S. (Eds.). (1982). *An international perspective on the evaluation of written communication*. New York: Pergamon.

Silva, T. (1989). *A review of research on the evaluation of ESL writing*. ERIC Document Reproduction Service, ED 409 643.

Taylor, J. R. (1995). *Linguistic categorization: Prototypes in linguistic theory* (2nd ed.). Oxford: Oxford University Press.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*, 169-193.