Council of Writing Program Administrators

WPA-CompPile Research Bibliographies

Richard H. Haswell, Norbert Elliot

Holistic Scoring of Written Discourse To 1985
(WPA-CompPile Research Bibliographies, No. 27)

October 2019

We have created two versions of our bibliography of early publications on holistic essay scoring in the UK and the US. One version is the text you are reading, which annotates 22 research studies and 10 early handbooks—a selection of works especially influential in the years covered (up to and including 1985). The second version is unabridged, an ongoing bibliography covering the same time period. Not yet complete, probably never to be complete, but with completeness imagined as its end, this bibliography currently annotates over 1,000 entries. The entries are entered into CompPile, where its unique search term is "No. 27" in the Annotations field. They connect to our *Early Holistic Scoring of Writing: A Theory, a History, a Reflection* (Utah State University Press, October, 2019) in the sense that the book is based on the scholarship but could not possibly reference so many works. The 32 items annotated below will also appear in the CompPile bibliography, but note that the annotations in the two versions are not identical, and that one annotation will supplement the other.

When an open-access internet bibliography accompanies a printed book, authority is rendered contingent. And so it is that the bibliographies might usefully be thought of as intersections of what has come to pass in print and what will continue to unfold on the web.

Our book, this curtailed research bibliography, and the expanding storehouse of over 1,000 annotated entries are all based on the same definition of *holistic scoring of essays*. We stipulate such scoring as the use of a scale to assign a single value mark to a whole essay and not separately to separate aspects of the essay, with scorers trying to apply the scale consistently, and with the final score for each essay derived from two or more independent ratings. The phenomena captured in this definition of holistic scoring—whether called by that name or some other name such as "general impression" or "rapid impression"—held the interest of researchers and evaluators in the UK and US for two thirds of the twentieth century. The enthusiasm for holistic scoring may have reached a peak in the 1980s, but holistic scoring endures in classrooms and in large-scale assessments, where a direct sizing up of whole pieces of student writing, rapid and gestalt-like, is seen as more useful to educational stakeholders than indirect measures such as counting of textual features or correct answers to items in limited-response tests.

Try as we might, we find it difficult to identify any method of assessing written communication that historically has suffered so much critique, endured so valiantly, and yielded such valuable information—not only about student ability but also about how our profession of writing studies designs research, makes inferences, offers qualifications, and draws conclusions. The more we learned about the history of holistic scoring, the more we learned about our profession and its understanding of our students and the complexities of situated language use that binds us all.

*Cite as: Haswell, Richard H.; Norbert Elliot (October 2019). Holistic Scoring of Discourse to 1985, WPA-CompPile Research Bibliographies, No. 27. *WPA-CompPile Research Bibliographies*. https://wac.colostate.edu/docs/comppile/wpa/Haswell-Elliot.pdf. Date of access.

Attitudes Toward History

In the decades when holistic essay scoring was first taken up by UK and US educational researchers and assessors (1930s-1940s), organicism flourished in nearly all fields of thought, from anthropology to ecology, from learning theory to neurology. Most professional fields appropriated the terms *holism* and *holistic* soon after Jan Smuts offered them to the public in *Holism and Evolution* (1926). In an intellectual environment teeming with holism, holistic essay scoring may have been only one small, intermittently growing plant. But it growth was not simple. To understand it, maps are useful. Here we offer eight that help navigate our two bibliographies (the first seven maps are from *Early Holistic Scoring of Writing*, pp. 9-25).

- *Historical constancy*: The history of holistic scoring and the published commentary on it are embedded in larger, archetypical continuities: the rise and decline of discourse genres; the slow accretion of research methods and findings; and the shifts in formal evaluation systems. Attention to these constancies helps us understand why, over many decades, holistic scoring became connected so centrally to every other method of writing evaluation—and to the growing body of knowledge about written communication.
- *Emplotment*: As Hayden White famously proposed in *Metahistory: The Historical Imagination in Nineteenth-Century Europe* (1973), historical meaning derives from narrative choice. This constructivist view of history, important to our book, also helps interpret this research bibliography. As readers review our annotations, they might well consider if the commentary assembles history as record, story, trajectory, social exploration, or global journey.
- *Vignette*: A research study is a circumscribed happening, and so is a discursive annotation of its occurrence. Narrative descriptions such as those found in our research bibliography are useful as heuristic devices that allow us to identify trajectories, examine institutional histories, and complicate received historical views.
- *Evidence*: Should a research study, and an annotation of it, be narrative, descriptive, explanatory, or predictive? Should it be empirically qualitative or quantitative? Should it use a mixed or multiple methodology? The answer is yes. Such eclecticism can lead us to reimagine evidence and expand our targets of study to include a broad spectrum of evidence: archives, in-house reports, structured interviews, authorial experiences, and other forms of evidence used to document lived experience.
- *Terminology*: Our research bibliographies often employ a specialized vocabulary (see the Keywords to these entries, and the Glossary in *Early Holistic Scoring of Writing,* pp. 291-299). Usually the technical language follows usage from the mid-1930s to the mid-1980s in the UK and the US. Bound to its time, the definitions are interesting in and of themselves as evidence of semiosis and situated language use.
- *Ecology*: Entailed in the definition of holistic scoring above is the operational testing ecology that allows holistic scoring to function (Lucas, 1988, Part 1 and 2). That ecology includes topics given, time allowed for composing, types of rater scales, types of rubrics and scoring guides, use of anchor essays, kinds of rater training, calculations of final score from independent scores, and subsequent statistical validations of any and all of the testing apparatus. It may seem that the research literature gives undue attention to this ecology, sometimes employing highly technical statistical maneuvers, but the ecology

raises profoundly important questions of the way scores are interpreted and the way information and outcomes impact test takers.

- *Continuum*: Over the decades, several distinct essay-scoring genera took root, ranging from analytic to holistic. By attending to these different continua, we better understand international trends in construct validity, reliability evidence, and consequences of score use. The distinct practices evolved, passed on by researchers, test constructors, testing administrators, raters, and teachers—and in the process acquired detractors, defenders, and devoted followers. Historically the earliest of these procedural methods was *connoisseurship* (the term is from Weir, Vidakovic, & Galaczi, 2013, p. 208). An essay is graded or scored by a single reader, typically using an internalized academic scale, the mark unchecked by any other reader. By the first decades of the twentieth century, *sample matching* was popular, at least in school settings. An essay is assigned a score or rank by fitting it within a given set of essays arranged from best to worst (e.g., Boyd, 1926, below). During the same time *analytic scoring* was also popular in both the classroom and the examination auditorium. The essay is scored by means of a checklist of criteria, each rated on its own scale. Analytic scoring still remains popular, among teachers (see below Diederich, 1966) and researchers (e.g., see below Anderson, 1960, Percival, 1966, Nold and Freedman, 1977). Neither connoisseurship or analytic marking, of course, qualify as *holistic scoring*, at least by our definition. Historically, the earliest bona fide method we call holistic is *pooled-rater scoring* (e.g., see below Hartog, Rhodes, & Burt, 1936; Cast, 1939, 1940; Wiseman, 1949; Britton, Martin, & Rosen, 1966; Godshalk, Swineford, & Coffman, 1966). Readers receive no or slight training, controlling devices such as rubrics or anchor essays are seldom used, and independent scorings of an essay, usually three or more, no matter how discrepant, are simply averaged or summed for a final score. For decades in the US, the most popular holistic method was what we call *adjusted-rater scoring* (e.g., see below Diederich, 1946; White, 1973; Myers, 1980), in which scorer consistency is promoted through scoring guides, anchor essays, and in-depth scorer training—and in which undue discrepancy in independent scores is not tolerated but resolved by adjusting the score, perhaps with a table-leader reading. The method was adopted and spread through College Board scoring of Advanced Placed composition essays. In the 1980s, yet another tradition of holistic scoring gathered steam, which we call *trait-informed scoring* (e.g., Hilgers, 1980). Scorers attend only to a few, selected essay traits, such as support or logic. The traits and scales may be arranged in a grid (Bossone, 1969), and sometimes the traits are individually scored and then summed for an "overall" score (Diederich, 1966, below). Trait-informed scoring borders analytic scoring and sometimes crosses over to it, as is the case with "primary-trait scoring" (Mullis, 1976) or "focused holistic scoring (Texas Education Agency, 1980, Breland, 1983). In fact, all these different scoring procedures have been lumped under the single term "holistic"—a questionable step that risks losing sight of the fact that their different scoring procedures are undergirded by very different and often incompatible philosophical and statistical rationales.
- *Aesthetics*. Art historian Francis Halsall (*Journal of Art Historiography*, December 2012) has observed that a judgment of fit—of match between an entity and its description—is compatible with aesthetic judgment. "This is to say," he notes, "that there is an aesthetic judgment at work in this understanding of the 'rightness' of the 'fit' between a way of worldmaking and the world. In other words, we might judge a description of a work of

> art— just like a work of art itself—by how satisfactory or even how pleasing the 'fit' is between its structural composition and the world that it describes. Does it feel right?" (p. 12).

In our view, this "recourse to pleasure," as Halsall terms it, is not necessarily a turn against evidence or validation. Rather, it is an admission of the human act of value making. It pertains to our own pleasure, as authors, in working on this research bibliography. While undoubtedly no perfect, our work feels right in its alignment of our narrative with the historical events we sought to capture. We are hopeful that the online annotations and the ones provided here will be read, challenged, and supplemented by those who will come after us with their own ideas of truth—and beauty.

The Research Studies

We begin with the research studies. From *The Marks of Examiners* (1936) to *Writing: Trends across the Decade, 1974–84* (1986), UK and US studies are notable in their research designs and attention to evidence gathering. However named, holistic scoring is central to a rigorous evaluation method that played a significant role in educational measurement during its most formative period. Watching the use of holistic scoring in varied settings for varied research aims, we can glimpse the influence that a serious assessment genre had on the shaping of a profession.

Note that we order these twenty-two studies chronologically not by year of publication but by year the data was collected.

Hartog, Philip Joseph, Edmond Cecil Rhodes, Cyril L. Burt

The marks of examiners: Being a comparison of marks allotted to examination scripts by independent examiners and boards of examiners, together with a section on a viva voce examination

London: Macmillan (1936)

This volume contains the first large-scale investigation empirically examining a true holistic scoring of essays. In their preface to the volume, Sir Michael Ernest Sadler and Sir Philip Joseph Hartog establish the need for the research contained in it: "No element in the structure of our national education occupies at the present moment more public attention than our system of examinations. It guards the gates that lead from elementary education to intermediate and secondary education, from secondary education to the Universities, the professions, and many business careers, from the elementary and middle stages of professional education to professional life" (p. vii). The research studies recorded in the volume were the outcome of an International Conference on Examinations held in May of 1931, at the sea town of Eastbourne, sponsored by the Carnegie Corporation, the Carnegie Foundation, and the International Institute of Teachers College at Columbia University. Hartog and Rhodes studied school and college examination essays (called "scripts" in the UK). They included studies of School Certificate Examinations (taken annually by 60,000 and 70,000 students at age 16), College Entrance Scholarship Examinations, and University Honors Examinations in Mathematics and History. But it was for their study of Special Place Examinations (the "11-Plus" exams taken annually by 400,000 to

500,000 students at ages 10 to 12 for secondary school admission) that Hartog and Rhodes applied holistic scoring. Around 1934, they had ten experienced teachers score seventy-five essays by "impression" on a hundred-point scale. Teacher-raters used no pre-set marking scheme, and two or more independent scores of an essay were not adjusted but merely summed (*pooled-rater scoring*). The teachers then read another set of seventy-five essays, judged of equal quality and range, on a weighted analytic scale, scoring separately the criteria of ideas, vocabulary, grammar and punctuation, sentence structure, spelling, and handwriting (*analytical scoring*). The researchers compared the two methods of scoring, holistic and analytic. The general impression scores had more range and the analytic scores more lenience. Accompanying *Marks of Examiners* was memoranda by Cyril Burt that lent statistical support to the conclusions Hartog and Rhodes had drawn regarding reader reliability. Key is Burt's finding regarding the usefulness of general impression marking (p. 240). By pooling the general-impression scores, the study had shown one way to a genuine holistic method of essay scoring—a way to determine an essay's score that was more reliable than the score of a single rater. As Burt recognized, the total impression is the result of distinguishable elements (identified in the analytical marking scheme) that readers may value very differently (p. 249). Viewed as more tentative than conclusive in its treatment of validity and reliability, *The Marking of English Essays* is far from a statement of triumphalism. Even in the earliest studies when problems were found in the reliability of markings, researchers did not recommend an a priori solution, instead recommending further systematic experimentation in order to rectify "the distressing uncertainties of the present examination" (p. xviii).

KEYWORDS: analytic-holistic, 11-Plus, interrater-reliability, leniency, pooled-rater, range

Cast, B. M. D.

The efficiency of different methods of marking English compositions (Part 1). *British Journal of Educational Psychology* 9.3 (1939), 257–69

The efficiency of different methods of marking English compositions (Part 2). *British Journal of Educational Psychology* 10.1 (1940), 49–60

B. M. D. Cast opens her two-part study of scoring UK examination essays with a quotation from Philip Boswood Ballard (*The New Examiner*, University of London Press, 1923): "The modern examination is dominated by the essay. It is based on the essay, it is built of the essay, it stands or falls by the measurability of the essay" (Cast, p. 257). As Philip Joseph Hartog, Edmond Cecil Rhodes, and Cyril L. Burt well knew (above), the genre of the written essay examination had long been adopted, without question, as a valid means of testing. Nevertheless, as a test item that consistently had proven difficult to mark reliably, it should perhaps be "cast on the dust heap," just as Ballard recommended (Cast, p. 258). Cast—a student of psychology at University College, London—saw an opportunity for systematic inquiry using analysis of variance (ANOVA) techniques. Her sample was drawn from forty compositions written in 1934 by 40 schoolgirls14.5-to-15.5-years-old. The scripts were scored by 12 teachers, according to the method they most commonly used. As a result, 5 out of the 12 judges relied on general impression, 3 on detailed analysis, and 4 on a combination of both. So Cast does not investigate true holistic scoring, but rather compares analytic with impressionistic. In terms of success in differentiating between the candidates, Cast found the analytic method best, with general

impression of almost equal value (p. 268). So Cast's work proved influential with later evaluators who wanted to compare the virtues of analytic and holistic scoring. However in 1939-1940, the most venturesome part of Cast's study was the extension of correlation analysis from the essays to characteristics of the student writers themselves. To create this correlation between writing ability and personality, Cast first established standardization of scores related to the elements of style, vocabulary, subject matter mastery, logical arrangement, and mechanical accuracy. Then, positive and negative correlations—termed saturations—were established between the elements. Based on a table titled "Marks for Different Aspects of English Composition" (p. 56), Cast then drew this inference: "Thus those with positive saturations include the more unstable or extraverted children, those with negative saturations are either highly stable or (more commonly) repressed extraverts. The former are imaginative, well informed, fluent writers, usually (though by no means always) fairly logical in arranging their ideas, but decidedly hurried in their writing and careless in their spelling and grammar. The latter are slow and careful writers, with little interest in external subjects as such and comparatively devoid of verbal fluency and elasticity of phrase. We might perhaps call them the fluent and the precise type respectively" (p. 57). In connecting performance to personality in terms of conceptually related constructs, Cast was among the first scholars to propose relationships between cognitive and noncognitive assessments.

Keywords: analytic-holistic, ANOVA, criteria, discrimination, essay-scoring, general-impression, personality assessment

Wiseman, Stephen

The marking of English composition in grammar school selection

British Journal of Education Psychology 19.3 (1949), 200–209

With no fanfare and for years no notice, R. K. Robertson, Chief Examiner for County Devon in the west of England, put into place the first genuinely holistic scoring of essays in an actual examination setting. In the spring of 1940 he had some 2,500 students taking the 11-Plus examination in English write an essay. They had previously scored in the middle third of the standard short-answer test. Each essay was scored on a 13-point scale independently by four teacher-raters. The teachers read quickly and their scores were simply averaged (*pooled-rater holistic scoring*). Robertson's system was continued for nine years in Devon. Then Stephen Wiseman, who participated from the beginning and who had assumed the position of Chief Examiner in 1947, published a research study describing and validating the system. It was lucid, highly knowledgeable, and broadly influential. Wiseman established test validity in terms of intrarater reliability. "The consistency thus being measured is consistency with some supra-individual standard, and it is arguable that this is as much a coefficient of validity as of reliability" (p. 203). General impression marking, he found, was superior to analytic marking, based on an explicit model of the writing construct (spelling, sense, punctuation, vocabulary, power of expression, and grammar). To treat student writing in such a way was, by association, evidence of fairness. As well, Wiseman emphasized the importance of criterion measures as they are related to scores: "A validity study for 11+ composition must, therefore, be a follow-up into grammar school achievement" (p. 205) (cf., Valentine & Emmett, 1932; Huddleston, 1948;

Nisbet, 1955). In his evidence gathering, Wiseman's position is clear: "The writer [Wiseman] is a confirmed 'general impressionist' having frequently had the experience of marking school essays by analytic methods (which for teaching purposes have obvious advantages) and finding that the obvious 'best' essay is not at the top of the list, the total gestalt is more than the sum of the parts. When we are faced with the task of judging any complex psychological material, it is probable that the method of total impression (provided the observer is suitably orientated) will yield sounder judgments than will analytic methods" (p. 205). This passage contains an early use of the term *gestalt* as an explicit classification of impressionistic scoring. (For an earlier use, see Cyril Burt, 1917, p. 27. We note, however, that in Gestalt theory of perception the whole is *other* than, not more than, the sum of its parts.) Wiseman's passage is equally remarkable because it contains one of the earliest distinctions between selection of scoring methods and score use. For assessment purposes, impressionistic scoring—compelling the reader to focus, in a principled fashion, on the gestalt of the student work—is ideal. However, for instructional purposes, use of identified traits—inviting readers to provide judgment on an identified model of writing—was superior because the elements of the model provided focus in the classroom. Each method had its use, depending on the assessment aim. In this distinction, Wiseman's study is forward-looking in connecting method to aim. Wiseman emphasized that the scoring methodology was the creation of Robertson and hoped that the 1949 study would "bring some (belated) recognition of his work" (p. 205).

KEYWORDS: interrater-reliability, holistic-scoring, intrarater-reliability, grammar, 11-Plus, data, general impression, analytic-holistic, pooled-rater, Devon, R. K. Robertson, Blitz, intrarater-reliability

Diederich, Paul B.

Measurement of skill in writing

School Review 54.10 (1946), 584-592

This is the first published description of formal holistic scoring in the US as applied in operational testing situations. Although in this piece Diederich recommends good ways for teachers to grade the papers of their students, in doing so he also indicates the method of grading essays written by undergraduates in The College of the University of Chicago, beginning in 1943. The scoring was for course credit and under the direction of the Board of Examinations, where Diederich was employed. Diederich first notes two sources of inconsistency in grading or scoring: interrater reliability (which he calls "objectivity"), and writer reliability or changes in quality over time in writing produced by the same student ("reliability"). He reports a study in 1943 that found one-fourth of University of Chicago students changing their scores when writing a second essay three weeks later (p. 587). Essays written on different topics will "never attain a correlation higher than .55" (p. 587). Consequently, for the writing examination at the University of Chicago, students were given six hours to write two essays. As for scoring, two independent readings were essential, by scorers who agreed upon criteria and scale: "If two readers are reading the same papers, let them look for the same things and let them mark on the same scale. Then their differences will be brought out, and the students will be protected against individual quirks of judgment" (p. 590). According to Diederich, readers should undergo some training: "after every reader has marked the same four or five papers and compared results, the marks will

come closer together, but there will still be discrepancies" (p. 589). In formal scoring, if there is a difference in score, "readers should re-read the paper together, explain the basis of the mark originally assigned, and agree on a reasoned compromise." If scorers "have come to agree fairly well on common standards, only about one paper in every ten or twelve will have to be examined [adjusted]" (p. 589). From internal Board of Examiners memoranda written by Diederich, we know that his system was applied and studied at the University of Chicago from 1943-1949. It qualifies as a true *adjusted-rater* form of holistic scoring: training of raters in criteria and scale, insistence on two independent global ratings of each paper, mutual consultation of raters to resolve discrepant scores (see *Early Holistic Scoring,* pp. 99-105, 132-137).

Keywords: assessment, evaluation, measurement, reliability, skill-level, University of Chicago, Basic, College, analytic scale, data, writer-reliability, interrater-reliability

Coward, Ann F.

The method of reading the Foreign Service Examination in English composition. ETS RB-50–57

Princeton, NJ: Educational Testing Service (1950)

In the US, Ann F. Coward at the Educational Testing Service performed the first comparative study of analytic and holistic essay scoring, or what she calls “atomistic” scoring and “wholistic scoring.” (This may be the first time the term “holistic” was published in connection with essay scoring; her spelling, however, was not unusual in 1950. Note that “holistic” as an essay-scoring term is an Americanism and never caught on in the UK.) Using 100 three-hour essay examinations from the 1949 fourth General Examination for the Foreign Service, Coward (with the assistance of psychometrician Frederick M. Lord), had four distinct tasks scored by both methods. “Wholistic” scoring was a "subjective, intuitive, over-all judgment of a composition," rating an essay on a scale of 1 to 10 on "total merit," judging "boldly, decisively, and rapidly" (p. p. 3). For each essay, four independent scores were summed (*pooled-rater scoring*). Two findings of the Coward study are notable. Of the four tasks scored in the study, the third (on the basis of a given statistical table, the candidate was asked to write an informative report of 300 words) yielded the lowest interrater correlations (0.44 to 0.63). Because this was not a typical writing task for an essay, it appeared that interrater reliability may vary according to task. Second, the analysis was subjected to differing statistic treatments that, in turn, led to different results. Those who read “wholistically” were given strict scoring rules: to place at least one paper in the lowest and one in the highest category, to not place more than 10 papers in either extreme category, and to use all categories, with a peak near the middle of the 10-point scale. Although there was no evidence that the 100 candidates formed a Gaussian distribution—indeed, the candidates for the Foreign Service examinations were likely to be verbally skilled and would, hence, demonstrate a left skewed curve with the higher scores on the right of the x axis—the instructions forced a normal distribution. Since no such instructions were given to those who scored atomistically, the correlations with the holistic scoring may be understood as an artifact of the scoring instructions. In the US, as indicated by this early study, the relationship between scoring and task was understood to be complex, as was the relationship between reader instructions and the inferences that would later be drawn from their scores.

KEYWORDS: assessment, evaluation, measurement, reliability, skill-level, University of Chicago, Basic, College, analytic scale, data, writer-reliability, analytic-holistic, “atomistic scoring”, Educational Testing Service, Foreign Service Examinations in English, interrater-reliability, leniency, task, “wholistic”, validity

Finlayson, Douglas Scott

The reliability of the marking of essays

British Journal of Educational Psychology 21.2 (1951), 126–134

Douglas S. Finlayson, a student pursuing the Bachelor of Education degree at the University of Edinburgh when he performed this study, advanced the emerging UK program of research by making inferences based on multiple forms of evidence, with particular attention to interrater reliability, intrarater reliability, writer reliability, and validity. His is the earliest study to use such a wide range of evidential forms. In his study, two parallel sets of four topics were given to the same group of children, with each child allowed to choose the topic for each essay. All children (N = 850) were enrolled in their final primary school year (with a mean age of 12.19-years-old) in 21 Edinburgh schools in different parts of the city. A random sample of 197 children were selected for the study. General impression scoring was applied in the study. Using analysis of variance (ANOVA) techniques, Finlayson found that a team of four markers could be expected to have a mark re-mark correlation of .94 (marks pooled). Mean intrarater reliability for one marker was estimated to be .691, and that for a team of six markers could be estimated to be .86. Where idiosyncrasies of markers as well as the day-to-day fluctuations of children are taken into account, the over-all reliability of the essay for a team of three examiners was estimated to be .79. If the children had written two essays instead of one, the reliability could be expected to rise to .88. In Finlayson’s research, we see one of the earliest links of writing ability (assessed through a direct measure) to criterion variables of IQ (Intelligence Quotient) and EQ (English Quotient), both measured through limited response, objective items. These measurement tools were developed under Godfrey Thompson at the Moray House School of Education. In the correlations, Finlayson found evidence that different constructs were being measured and that writing was, perhaps, a construct unto itself.

KEYWORDS: intrarater-reliability, interrater-reliability, intrarater-reliability, writer-reliability, direct-indirect, pooled-rater, holistic, Wiseman, replication, data, test-retest, cost-analysis, ANOVA, English Quotient, general impression marking, Intelligence Quotient, variability in marking, test-retest reliability

Anderson, C.C.

The new STEP Test as a measure of composition ability

Educational and Psychological Measurement 20.1 (1960), 95–102

As is the case with Finlayson, Anderson of the University of Alberta also uses analysis of variance (ANOVA) techniques—in this case, to examine essays written for the Sequential Tests of Educational Progress (STEP), developed by the Educational Testing Service in 1957. While

the STEP Writing Test constituted only multiple-choice questions, to be used from grade 4 through the sophomore year of college in order to establish continuity of measurement for individual students, there was an accompanying essay test that could also be used. Focusing only on written compositions, Anderson elicited 8 STEP essays from 55 students. They were selected according to an Intelligence Quotient (IQ) range of 100, with a standard deviation of 13—a level of average intelligence. Three markers scored the 440 essays on a seven-point scale in which quality of thought was worth 50 percent of the final mark, style 30 percent, and sentence structure 30 percent. The scale had been developed by Paul Diederich to be used with STEP essays. Since Anderson, researchers have been divided in calling his scoring method analytic or holistic. Analysis of variance revealed statistically significant interaction effects among testing occasion and marker. As well, students themselves varied across the writing tasks: Of the 55 students, 71 per cent received different scores on different tasks (compare Diederich, 1946, above). As Anderson concludes, "The marking schedule of the new STEP Essay Test has not reduced, at least in this experiment, into insignificance the extent of variability characteristic of well-known sources in the marking of essays-composition fluctuation, the unrepresentativeness of essay samples, and discrepancies among markers. Of the 55 students who wrote essays, 71 per cent showed evidence of composition fluctuation and 78 per cent stimulated discrepancies in marking among markers. Fifty-six per cent showed evidence of both" (p. 101). Analysis of variance techniques had explicitly revealed, in quasi-holistic scoring, multiple sources of variance related to reliability, with special emphasis on construct representation as manifested within tasks. Later researchers paid close attention to Anderson's study, and would delve much more deeply into the effects of writing task (time and topic) on holistic scores (e.g., Gray, et al., 1982, below).

KEYWORDS: ANOVA, Intelligence Quotient, Sequential Tests of Educational Progress, STEP, measurement, direct, validity, reliability, pooled-rater, instability, interrater-reliability, writer-reliability, error, reliability, validity

Diederich, Paul B., John W. French, and Sydell Carlton

Factors in Judgments of Writing Ability. Research Bulletin RB-61–15

Princeton, NJ: Educational Testing Service (1961)

For US researcher Paul B. Diederich and his ETS colleagues, two sources of variation—student performance as related to task and construct underrepresentation in the task—were not of immediate interest in the way that Anderson (above) had formed the problem. Rather than conduct a study and then make inferences on what the possible problems might be in drawing inferences, Diederich—with psychometrician John W. French and statistician Sydell T. Carlton—decided on a bottom-up approach. The study, conducted in 1959, employed the fairly novel method of factor analysis. Based on matrix algebra, the method is described in *Vectors of Mind: Multiple-Factor Analysis for the Isolation of Primary Traits* by L. L. Thurstone (University of Chicago Press,1935) of the University of Chicago. Using matrix theory, Thurstone held that information—in this case, a score—can be expressed as a linear function of a number of factors, not as a single factor. The multi-factorial methodology was an expression of Thurstone's desire to create "generalization of the factor problem to *n* dimension" (1935, vii)—a

mathematical solution to what he described as the faith of all science: "that an unlimited number of phenomena can be comprehended in terms of a limited number of concepts or ideal constructs." "Without this faith," he added, "no science could ever have any motivation" (44). Following Thurstone, the Diederich team hypothesized that if the factors of writing upon which readers make judgments could be identified, then perhaps the variables of writing could be identified. This method yielded five principle factors that the researchers named ideas, form, flavor, mechanics, and wording. A construct model was thus, for the first time in history, empirically derived from the comments of readers responding directly to student written texts. Implicitly, if validity evidence was to be demonstrated, it could be collected as the five-factor model was used. However, in the larger history of holistic scoring, the creation of the factor model was of little historical consequence—although it played a major role in scaled analytical scoring (e.g., Diederich, 1966 below), and computer scoring (e.g., Page &, 1966, 1968). In our interpretation of the study, the key finding appears on p. 33: sixteen of the fifty-three readers had loadings on the first factor of 0.25—more than any other factor. Further details of the factor analysis, placed in Table A-3 (p. A9), showed the factor correlations. We quote the report here: "It is by inspecting this table that we can learn something about the 'general factor' of reader agreement which was discounted by the particular kind of rotation of factors that was chosen. Inspection of the table reveals that we have a table of rank one. That is, if we were to factor analyze this table of intercorrelations, we would find only one 'second order' factor. The over-all agreement of the readers and tests that is not explained by the six factors already discussed appears to be concentrated on one rather than several major aspects of writing" (p. A6). So there it was, empirically verified by state-of-the-art methods in which multiple phenomena were reduced to one: the holistic score.

KEYWORDS: arrangement, CEEB, commenting, criteria, diction, factor analysis, flavor, general merit, holistic, ideas, form, MX, diction, content, grading, scale, criteria, factor-analysis, interrater reliability, mechanics, pooled-rater, style, vocabulary

Diederich, Paul B.

How to measure growth in writing ability

English Journal 55.4 (1966), 435–449

After the 1961 publication of *Factors of Judgements of Writing Ability*, Diederich took to the pages of *The English Journal* to present the study results to classroom teachers. A brief history of the factor analysis study is provided, but the construct model is somewhat different from that empirically derived, with mechanics expanded to include a list of errors not found in the original report, including usage, sentence structure, punctuation, capitals, abbreviations, numbers, spelling, handwriting, and neatness. Here Diederich is among the first U.S. researchers to take research into the field of teaching. As Sydell Carlton told us in our 2014 interview with her, her colleague was devoted more to "making students' lives richer." The five-factor model (expanded to eight factors in the article) is presented by which, Diederich claims, growth in writing ability can be examined across school years. Based on a method of sampling across grades, he claims, "[Y]ou can get a solid and convincing answer to that question [how much growth in writing ability comes about in each year of your program] in a single weekend, starting Friday morning

and reporting results at the close of school on Monday" (p. 435). In terms of scoring, teachers simply sort the randomly selected papers across grades into three piles of high, medium, and low. The differences across grades, he reminds readers, will not be "sufficiently reliable for individual measurement" but will nevertheless be "reliable enough to measure the difference between one grade and the next" (p. 437). His was among the earliest studies to bring the concept of academic program validation, based on scores from writing samples, into the literature of direct writing assessment. This 1966 study is the first mention of what would become known as the Diederich scale—one of the most influential and popular methods of direct writing assessment, and an early example of the common choice by teachers and researchers of analytical breakdown over holistic scores. The popularity of the Diederich scale was further spread through his handbook *Measuring Growth in English*, NCTE, 1974 (see below).

KEYWORDS: analytic-holistic, Diederich scale, growth, program-evaluation, assessment, measurement, scale, ranking, improvement, school, curriculum

Godshalk, Fred I., Frances Swineford, William E. Coffman

The measurement of writing ability

New York: College Entrance Examination Board (1966)

Edward S. Noyes, a professor of English at Yale University and former director of admissions there, was ecstatic in his introduction to *The Measurement of Writing Ability*. "It is clear from this monograph," he gushed, "that colleges can in general accept scores on the English Composition Test in any of its current forms as valid indices of their candidates' ability to write (p. vi). Fred I. Godshalk, Frances Swineford, and William E. Coffman were more introspective about the findings of their report. They had read Philip E. Vernon's *Secondary School Selection* (Methuen, 1957) and referenced the very pages (pp. 120-121) in which studies by Hartog, Rhodes, Burt, and Wiseman were discussed in detail (see above). One imagines the ETS researchers were drawn to those pages both because of Vernon's summary and because of his advocacy of "general impression marking" and the use of "total pooled marks" as strategies by which "the inherent subjectivity of essay-marking can be reduced to reasonable proportions, in examinations where the range of pupil ability is wide" (p. 121). As Godshalk, Swineford, and Coffman write, their study "supports conclusions" already reached by Vernon (p. 39). In this light, the 1966 published study may be usefully situated as a replication of previous U.K. studies. It is equally important to establish that the U.S. research team was also influenced by the larger university community. On October 2, 1961, Swineford and Godshalk first proposed the idea for a validation study of the English Composition Test ("Memorandum for All Concerned," Princeton: Educational Testing Service). They referenced Osmond E. Palmer, who was Examiner in English at Michigan State University and chair of the CEEB English Composition Committee. To validate the interlinear exercise in the ECT, Palmer insisted on a criterion variable of five essays by each student writer and five independent readers for each essay, stipulating that "The readings would be holistic, rather than analytic, but they would be based upon a group examination and discussion of selected papers, with consensus reached as to the ratings of each paper and the elements that had been considered in arriving at the decision" ("Memorandum," pp. 2-3). As Godshalk, Swineford, and Coffman's final 1966 report notes, by using five raters per essay, five

essays per student, and a 3-level scale, the researchers found that rapid holistic scoring of essays is reliable and adds unique information to objective testing and interlinear exercises. They then draw the subsequent inference: “An essay in the English Composition Test says to the student that skill in actual writing is an important outcome of education. It says to the teacher that the ability to answer multiple-choice questions, unless accompanied by the ability to compose answers to essay questions, is not sufficient evidence of effective teaching” (1966, p. 41). In taking a direct measure of writing as the criterion against which the other measures were validated, Godshalk, Swineford, and Coffman located the writing sample in the center of U.S. educational measurement of written communication. In the US, *The Measurement of Writing Ability* became one of the most influential reports of pooled-rater holistic scoring, in part because of the meticulous design of the studies included and the lucidity and passion by which they were reported.

KEYWORDS: evaluation, measurement, proficiency, direct-indirect, holistic, research-method, validity, data, pooled-rater, analytic-holistic, cost-analysis, interrater-reliability, validity

Myers, Albert E., Carolyn B. McConville, William E. Coffman

Simplex structure in the grading of essay tests

Educational and Psychological Measurement 26 (1966), 41–54

The question of research aim becomes important as historians of writing assessment shift back and forth between the United Kingdom to the United States in the 1960s. In the Britton, Martin, and Rosen 1966 experiment (see below), a study of 500 writing samples was undertaken to support UK classroom and testing practice. In the Myers, McConville, and Coffman study published the same year, 80,000 essays were examined to determine interrater reliability for large scale US testing. In the UK study, holistic scoring was viewed as a way to support the pedagogy of classroom teachers, in the US, holistic scoring was viewed as a way to manage large-scale assessments. Generally speaking, UK assessments efforts in the 1960s were often deeply contextual, and US studies were deeply formal. For Myers, McConville, and Coffman, the occasion for the analysis was significant. In December of 1963, some 80,000 students electing to take the English Composition Test (ECT) in the College Entrance Examination Board test administration were required to write a 20-minute essay. The test was the first ECT to include a 20-minute essay since April of 1947. The research therefore was a milestone moment of reflection in a sixteen-year period during which the College Board had shifted from an essay examination through various stages of objective and semi-objective format then back to an essay format. Of interest is the team’s identification of a new interpretative element related to factor analysis, an element that infers holism. Subsequent factor analysis of covariances among the 25 papers yielded 4 factors. As had been the case in *Factors in Judgments of Writing Ability* (1961), the Meyers team demonstrated that the factor loadings (4 in this case) were indeed giving global judgments. However, the Meyers team also observed that there was a functional relationship between scores and the factors, leading the researchers to conclude that there was a simplex structural relationship at hand. In 1954 Louis Guttman described a *simplex* as a structural relationship that exists between elements when these elements are ordered on a single dimension. Using this interpretative framework, the Meyers team hypothesized that essays may be classified

as complex in the sense that they embody the existence of varied attributes, from good organizational structure to command of conventions. An essay receiving low scores, conversely, is simple in that these attributes may not be present. "Thus," they conclude "the ordering of papers by their mean could be construed as an ordering by complexity" (p. 52). Thus, while the holistic interpretation was maintained, a nuance was provided: When scores were understood in terms of quality, diverse points of view were, perhaps, also part of the assigned score—with high scores based on observations of complexity in the writing sample and low scores based on observations of simplicity. Robust construct manifestation was, at least in theory, related to score level.

KEYWORDS: CEEB, English Composition Test, factor-analysis, simplex structure analysis, evaluation, essay-scoring, data, measurement, interrater-reliability, criteria, high-low, holistic, data, Educational Testing Service, pooled-rater, factor-analysis, halo-effect

Britton, James N., Nancy C. Martin, Harold Rosen

Multiple marking of English compositions: An account of an experiment. Schools Council Examinations Bulletin No. 12

London: Her Majesty's Stationery Office (1966)

In the mid-1960s in the UK, James Britton and his colleagues had hoped to provide empirical support for the classroom practice "of continuous writing . . . as a major part of what a pupil does in his English schoolwork, and that the form of the examination given to him ought to encourage rather than discourse such practice" (p. 3). In using 500 writing samples written in 1964 by 16-and-17-year-old British students from the General Certificate of Education (GCE) Ordinary Level (O-Level) Examinations, the researchers aimed to ensure that claims could be made about the role of writing in a family of basic academic qualification examinations used in England, Wales, and other Commonwealth nations. In replicating the marking system used by the GCE, the Britton team assured that their experiment would have a criterion measure of comparison for the rapid-impression, multiply-scored measure they promoted (and borrowed essential from Wiseman, 1949, above). In their experiment, a student's score was the total of three independent holistic scores on a 20-point scale (*pooled-rater scoring*). Written instructions to impression markers were kept as short as possible and there was no other form of "briefing, consultation or moderation" (p. 1). Pooled rates of three holistic markers had a reliability (.77) that surpassed the reliability of two Cambridge Board markers scoring in their usual analytical way. Independent judgments of essay qualities were more in agreement with the holistic marks. The cost of holistic, pooled-rater scoring was minimal. It is no surprise when Britton and his colleagues write that the "the actual work of students must be viewed in opposition to the use of objective tests." If "the practice of continuous writing ought, on the strongest educational recommendation, to be a major part of what a pupil does in his English school work," then it follows that "the form of the examination given to him ought to encourage rather than discourage such practice" (p. 3). The advice was accompanied by a warning about consequences of test use: If the government is to continue using limited means for assessing complex linguistic abilities, then it must "take responsibility of the 'backwash effect" of its actions (p. 3). As a result of the design—accompanied by detailed and rigorous analysis, safeguards, conclusions, and

implications of test use—*Multiple Marking of English Compositions* should be viewed as the most convincing empirical backing for pooled-rater holistic scoring of essays in large-scale examinations existing at the time.

KEYWORDS: analytic-holistic, subjectivity, experiment, interrater-reliability, grading, evaluation, data, holistic, general impression, pooled-rater student-opinion, 'multiple marking', washback ['backwash', p. 3], Wiseman, subjectivity, analytic-holistic

Percival, E.

The dimensions of ability in English composition

Education Review 18.3 (1966), 205-212

As was the case with Finlaysonn (see above), Percival turned to the UK Moray House tests for criterion variables. Percival investigated writing samples of an 11-Plus group of students enrolled in Bolter Grammar School in the English county borough of Warrington. Based on low to moderate correlations among measures, he determined that taking a measure of IQ was not taking a measure of writing ability. Similarly, absence of high levels of correlations between EQ ("English Quotient") and defined traits of writing suggested that objective tests were an inappropriate measure of writing ability. Percival extends the Finlayson study through a reflection on resonances between general impression scoring (holistic) and the schedule method (analytic scoring). Percival examined the relationship between the general-impression score and the selected trait combinations. The relationship between a combination of fluency, effective language, accuracy of composition, and accuracy of spelling was 0.93 for boys and 0.81 for girls. It was not that general-impression scoring was without foundation. Rather, it was that the study had led to a way of systematic identification of the "elements that the marker is most affected by when he makes his assessment by the general impression method" (p. 211). As was the case with the Diederich, French, & Carlton, 1961 (above), Percival was more interested in independent measure reliability concerning traits and general-impression scores than in interrater reliability. Fearful of backwash effects if the inferences were misinterpreted, Percival was careful to caution readers that the identification of elements did not mean that writing instruction was to be made elemental. "Miscellaneous exercises in punctuation, sentence formation, choice of vocabulary, and so on," he wrote, "will not alone produce skill in composition. The final product calls for the integration of these elements, not merely their sticking together. They are ultimately interwoven and they thus form a highly complex product. Part of the skill of teaching composition lies in this action of integration" (p. 212). Historically, Percival provided one of the first connections between instruction and assessment that emphasized the need for an integration of traits. In addition, his emphasis on complexity supports the view that composition is a unique event which cannot be simplistically dismantled into elemental parts.

KEYWORDS: 11-Plus, English Quotient, general impression, integration of traits, Intelligence Quotient, schedule marking, evaluation, criteria, direct, data, analytic-holistic, holism, essay-length, sentence-length, punctuation, singularity

White, Edward M.

Comparison and contrast: The 1973 California State University and Colleges English Equivalency Examination

Los Angeles, CA: Office of the Chancellor, California State University and Colleges (1973)

In the late spring of 1972, the Chancellor's Office of the California State University and Colleges (CSUC) agreed to support a summer study undertaken by a committee of the California English Council to investigate equivalency testing in English in the nineteen-campus system. The first English Equivalency Examination was given in late spring of 1973. *Comparison and Contrast* (1973) details the procedures and outcomes of that examination. When the report was published in October of 1973, Edward M. White and his co-authors provided a very influential model of a large-scale university-system assessment using holistic scoring with the aim of accurate advance credit and, if warranted, earlier graduation. With classroom instructors at California State University as readers, White directed the examination of 4,071 students who had taken a 90-minute multiple-choice College Level Examination Program (CLEP) test on literature and who had provided two 90-minute essays based on Advanced Placement English examination models. Two people versed in AP readings, Gerhard Friedrich and Rex Burbank, ran the scoring sessions and wrote detailed descriptions of them in White's report. The report quotes from Godshalk, Swineford, & Coffman, *The Measurement of Writing Ability* (1966): "The combination of objective items (which measure accurately some skills involved in writing) with an essay (which measured directly if somewhat less accurately, the writing itself) proved to be more valid than either type of item alone" (*Comparison and Contrast*, p. 108). White's report reveals a very fluid relationship between large-scale non-profit assessment organizations and locally-based assessments. In that relationship, there is resonance between testing and research undertaken at educational measurement non-profit organizations (in this case, AP scoring practices and Educational Testing Service researching for the College Board) and the impulse for localism (expressed on CSUC campuses). For more on that dynamic fluidity, see Godshalk, Swineford, & Coffman, 1966, above, and Haswell & Elliot, 2017. On the CSUC English Equivalency Examination in particular, see Haswell & Elliot, 2017, which analyzes that fluidity in detail.

KEYWORDS: holistic, direct-indirect, Advanced Placement model, College Level Examination Program (CLEP), Educational Testing Service, testing, equivalency, advance placement, California State University and Colleges, Freshman English Equivalency Examination, data, cost, budget, prompt, rating, follow-up, race, norming, uneven, Gerhard Friedrich, Rex Burbank, "key" (p. 42), "rubric" (p. 36), essay-scoring, adjusted-rater scoring

Nold, Ellen W., Sarah W. Freedman

An analysis of readers' responses to essays

*Research in the Teaching of English*11.2 (1977), 164–174

In a study completed in 1974, US researchers Ellen W. Nold and Sarah W. Freedman used multivariate analysis to tease out which textual elements of two timed writing samples written by

22 Stanford first-year students most influenced holistic reader response. The dependent variable was the holistic score on the two essays, teachers using a 4-point scale. There was a .85 correlation between scorers. The independent variable was based on four textual elements: the number, development and logic of ideas; the presence and appropriateness of the organization of those ideas; the complexity, variation and appropriateness of the syntax; and the richness and appropriateness of vocabulary. Knowledge of conventions associated with mechanical and spelling errors were not part of the analysis because Stanford students, the authors note, make few such errors. Measurement of the variables was established by counts of various kinds—words per T-unit, for instance, and categorization of finite verbs as common or not. Following stepwise multiple regression using R^2 as coefficient of determination, 20 percent of the variance was explained by length of essay in words, 3 percent by final ssentence modifiers (the authors incorrectly report 12 percent in the text for $R^2 = .312$), 4 percent by modal verbs, 4 percent by "be" or "have" as auxiliary modifiers, and 4 percent by common verbs. Their study found evidence that that length of essay was by far the single best predictor of holistic quality scores, a finding supported by later researchers. In another finding supported by later research, they found that college teachers were not very attentive to syntax: "Words per T-unit and other standard developmental measures are not useful in predicting perceptions of quality on the college level" (p. 174) (compare Freedman, 1979, below). While Nold and Freedman make clear that they have identified significant predictors of holistic scores, they were very alert to the tenuousness of their findings. "Care must be taken" they caution, "that both readers and tasks remain as consistent as possible across studies" so that "the research, the test maker, and the composition teacher" may benefit by the discovery of what textual elements most impact reader judgments" (p. 174). Still, their study was one of the earliest that critiqued holistic scoring by plumbing the textual features of essays that were associated, and not associated, with holistic scoring of those essays.

KEYWORDS: modification, stepwise multiple regression, essay-length, evaluation, reader-response, data, t-unit, MX, modification, free-modification, essay-length, multiple regression-analysis, reliability, Diederich scale, semantic, arrangement, syntactic complexity, vocabulary, measurement, syntax, active-passive, predictive, holistic, general-impression

Breland, Hunter

Group comparisons for the Test of Standard Written English. CEEB Research and Development Report 77–78, No. 1

Princeton, NJ: Educational Testing Service (1977)

Introduced in 1974, the Test of Standard Written English (TSWE) was part of the College Entrance Examination Board's Admission Testing Program. The 30-minute multiple-choice examination was one of 14 academic-subject tests designed to accompany the Scholastic Aptitude Test, so that college administrators could have information upon which to make admission and placement decisions from a single source. The Breland study was among the first to report group differences in scores based on a given construct of writing. In 1976, Hunter Breland and Gail H. Ironson had designed a comparative analysis of admission strategies following the *DeFunis v.Odegaard* decision that brought the issue of preferential minority admissions to the Supreme Court (*Journal of Educational Measurement*, 13.4, pp. 89-99). The

next year, using psychometric models associated with the determination of fairness as developed by Anne T. Cleary ("Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges," *Journal of Educational Measurement* 5.2, 1968,115–124) and Nancy Cole ("Bias in Selection," *Journal of Educational Measurement* 10.4, 1973, 237–255), Breland analyzed the TSWE in terms of differences between majority and minority groups on predictors and criteria, correlations between predictors and criteria, and slopes of the group regression lines. The results of regression analyses comparing TSWE scores and writing samples were especially revealing. The difference in the intercepts of the regression line between men and women was not statistically significant ($p = 0.057$)—but only barely so. Conversely, the differences between majority and minority students were quite statistically significant ($p < 0.001$)—a high statistical significance indeed. But for a rounding decision upward for 0.057 rather than downward of 0.05 (the borderline for statistical significance), differential prediction would have been established for both groups. Even in the most generous interpretation, it is difficult today to agree with Breland's conclusion that no important group differences were observed. Yet, while we may debate the conclusions drawn from the Breland study, we must also acknowledge that it paved the way for future studies. "The study," Breland concludes, "was limited to some degree, of course, by the necessity of combining all minorities into one group. Future studies should attempt to focus on single groups. Therefore, sufficient quantities of data should be collected for within-group analysis" (p. 44). In 1981 Edward M. White and Leon L. Thomas would do just that in "Racial Minorities and Writing Skills Assessment in the California State University and Colleges" (see below).

KEYWORDS: analytic-holistic, admission testing, bias, legality, CEEB, group comparisons, Test of Standard Written English, minority, group comparison, gender-bias, regression-analysis, data

Freedman, Sarah W.

How characteristics of student essays influence teachers' evaluations

Journal of Educational Psychology 71.3 (1979), 328–338

In a mixed-guise study done for her 1977 Stanford dissertation, Freedman had Stanford sophomores rewrite 32 student papers to be strong or weak in four different categories: content, organization, mechanics, and sentence structure. Teachers from Stanford's first-year writing program applied a 4-point holistic scale to all the rewritten essays. Essays written to be strong in organization received the highest average score, followed by content, mechanics, and sentence structure (the last with no effect). Of more interest is her exploration of interactions through analysis of variance, which revealed that it was the content variable that caused readers to score an essay significantly higher. Difference between the average score given papers weak in content and the average score given papers strong in content was 1.06 points, a large difference in relation to the 4-point scale. Conversely, effects of mechanics were about 1/2 of a point, and effects of sentence structure rewriting were approximately 1/4 of a point—although both were statistically significant ($p < .001$ and $p < .01$, respectively) in their correlation with organization. As in her 1977 study with Ellen W. Nold (above), Freedman closed her article with instructional implications: "If society values content and organization as much as the raters in this project and

many of the earlier studies apparently did, then according to the definitions of content and organization used in this study, a pedagogy for teaching writing should aim first to help students develop their ideas logically, being sensitive to the appropriate amount of explanation necessary for the audience" (p. 336). While she notes that many college-level curricula begin with a focus on mechanical and syntactic problems, it is important to supplement this approach with carefully planned curricula for teaching content and organization aspects of discourse. The vision was to be confirmed by George Hillocks, Jr.'s 1986 metanalytic study of different composition pedagogies (*Research on Written Composition: New Directions for Teaching*, ERIC Clearinghouse and National Conference on Research in English). The pedagogy with the largest effect size (0.44) was the environmental mode of instruction, with its attention to clear and specific objectives, problems selected to engage students with each other in defined processes important to particular features of writing, and activities facilitating high levels of peer interaction concerning specific tasks. Freedman's attention to the integrated nature of written communication was prescient in terms of knowledge yet to come about effective writing instruction.

KEYWORDS: organization, sentence structure, high-low, mixed-guise, arrangement, sentence, syntax, MX, evaluation, holistic, ANOVA, content, data, regression-analysis, measurement, rater-bias, teacher-rater, criteria, feature-analysis

White, Edward M., Leon L. Thomas

Racial minorities and writing skills assessment in the California State University and Colleges

College English 43.3 (1981), 276–283

Edward M. White and Leon Thomas present the first large-scale US account comparing score distributions in terms of minority status. The two score distributions were produced by indirect (limited response) and direct (holistic) methods. Information was presented on the performance of 10,719 first-time freshmen admitted to the California State University and Colleges in fall 1977 on both the Test of Standard Written English (TSWE) and the English Placement Test (EPT). The large sampling plan therefore served as the first large-scale study of a test of Standard Written English using multiple choice questions (an examination of what White and Thomas termed the grapholect) and a comparative assessment using both multiple-choice questions and holistic scores (described as a test of writing ability). While Breland, 1977 (above) had reported statistically significant majority-minority differences in a comparative regression analysis of essay scores on the TSWE, White and Thomas provided descriptive statistics in the form of histograms. While Breland had used inferential analysis, White and Thomas presented score distributions that had a straightforward message: direct measurement of writing produced different distributions for minority students than those for white students. For the overall group and for white students, White and Thomas reported a Gaussian distribution for the holistically scored essay, accompanied by a left skewed distribution for the EPT (that is, both the multiple-choice questions and the essay). The TSWE scores were flatter but nevertheless approached a normal distribution. For black students, however, the distribution patterns were described as "dramatic" (p. 280): the TSWE produced a right-skewed distribution, placing approximately 11 percent of the students at the lowest score range, while only the essay of the EPT had continued

to approximate a Gaussian distribution. Similar patterns were reported for Mexican-American students and for Asian-American students. As White and Thomas concluded, "The TSWE does not distribute minority students the same way trained and accurate evaluators of writing samples do" (p. 282). As a vehicle for construct representation, test type itself appeared to be related to the potential for disparate impact.

KEYWORDS: admission testing, test-bias, California State University and Colleges, California State University English Placement Test, group comparisons Test of Standard Written English, assessment, testing, racial, minority, TSWE, CEEB, SAT-testing, EPT, testing, direct-indirect, African-Am, Mexican-Am, Asian-Am, Anglo, topic, data, distribution, frequency

Woodworth, Patrick, Catharine Keech

The write occasion. Collaborative Research Study No. 1

Berkeley, CA: Bay Area Writing Project (1980)

In 1980, Patrick Woodworth, a California English Teacher at Tomales High School, and Catherine Keech, a research assistant at the Bay Area Writing Project, reported what was to become one of the first studies of the possible association of task design with holistic score. The study was straightforward. Students in three ninth-grade English classes and three mixed junior-senior level English classes wrote on topics in which no audience was identified, an imagined audience was suggested, or a real audience was identified. While the tasks were varied, they were aligned to classroom assignments in which students wrote about their own experiences. *Adjusted-rater* holistic scoring was used to evaluate student performance on all tasks. While grade-level differences in performance were found and women performed at higher levels than men, no statistically significant differences were found among the three different tasks. As Woodworth and Keech concluded, audience specification for the sample at hand "does not necessarily result in simple and direct improvement of student writing" (p. 34). As well, audience specification did not appear to result in genre shifts; the informal essay remained the default genre. Keech was especially candid about the questionable value of holistic scoring for identifying changes in student writing due to nuances in task specification: "Finally, we know that the method of holistic scoring generally registers only fairly large differences in writing performance between groups—differences created by such strong factors as age, socio-economic level, and sex. If the differences among groups caused by writing for different audience conditions were not major, they would be unlikely to affect mean holistic scores" (p. 25). Scoring shortcomings aside, the implications of the study were especially interesting. For students well adapted to school tasks, "the general audience of test readers can provide a real rhetorical context for some writers, and for some kinds of writing" (p. 34). At the classroom level, Woodworth interpreted this finding as evidence of the significance of "a sense of occasion" in which students are told that they are involved in an unusual writing situation—one in which "a world larger than the immediate class-room—in this case, readers/teachers from other schools—is involved" (p. 38). The sense of occasion, he speculated, may be useful in boosting motivation. This study is among the earliest US studies to vary task design systematically and examine the impact of group difference. It appeared as Part III of Gray, et al., 1982 (below), adding to that work's critique of task design in holistic scoring.

KEYWORDS: group comparison, holistic scoring, task design, school, teacher-cooperation, best-practices, assignment, sample, high-low, prompt-effect, topic-effect, audience specification, assessment, analytic-holistic, essay-quality, gender-difference, age-difference, data

Gray, James R., Leo R. Ruth (Eds.)

Properties of writing tasks: A study of alternative procedures for holistic writing assessment. Final report, NIE-G-80-0034

Berkeley, CA: Bay Area Writing Project (1982) [ERIC Document Reproduction Service, ED 230 576]

Along with the concurrent studies produced by the Center for the Study of Evaluation at UCLA (see Quellmalz, 1981, below), the research reported in this Bay Area Writing Project stands as the first sustained critique of holistic scoring. In 1980, BWAP won a National Institute of Education (NIE) grant to study the process through which writing proficiency and performance are assessed, in particular to investigate "the so-called 'holistic' approach to appraisal" (p. 1). James Gray, as Director of BAWP, was listed as principal investigator of the grant. As Project Director, Leo Ruth, professor education at Berkeley, led the study. Research coordinators were Sandra Murphy and Catharine Keech. Research associates were Karen Carroll, Charles Kinzer, Don Leu, Ann Robyns, and Elissa Warantz. Mary Ellen McNelly co-wrote one of the chapters. Sarah Freeman was consultant, and Marcia Farr liaison with NIE. James Gray and Leo Ruth sought to extend and deepen the research by Patrick Woodworth and Catherine Keech (1980, above) by differentiating between classroom writing assignments and school writing tests. In classroom contexts, Gray and Ruth noted, interpretation is possible, as well as preparatory help and motivational encouragement. In test conditions, however, no such contextualization is possible. "Each head bent over the page is presumed to be getting the same message to direct his/her writing performance," they noted (p. 34). Holistic scoring was used to tease out interactive relationships among participants, products, and processes during writing assessment episodes. Central to the analyses was a model of study of the writing assessment episode, with interactions among participants, processes, and products identified (p. 8). Task design, student response, and rating processes were the distinct and interrelated processes that allowed identification of the psycholinguistic and sociolinguistic variables involved in designing the writing assessment episode. In what was to become among the first systematic study of task response processes, Ruth and his colleagues were indeed able to identify variables associated with writing task design. Those variables encompassed the range of task design, from discourse mode (using topics that encourage introspection or autobiographical writing by drawing on personal experience) to task structure (using specifications to cue students to topic exploration and evaluator expectations). And, because agents were identified in the writing assessment episode and because response processes were used in the study, researchers were able to show that indeed student interpretations of the writing task differed from teacher and rater task interpretations. In determining the properties of writing tasks, the episodic model has proven invaluable to the present day in allowing granular study of design and impact. While the original 1982 report may long have gathered dust in its ERIC microfiche form, many readers were

influenced by the use Ruth and Murphy made of the report in their book *Designing Writing Tasks for the Assessment of Writing* (New Jersey: Ablex,1988).

KEYWORDS: discourse mode, task design, Writing Assessment Episode Model, data, experiment, holistic, topic, assessment, audience, interrater-reliability, interpretation, longitudinal, measurement, National Institute of Education, final-report, model, prompt, audience, rater response, student response, anchor, growth, development

Quellmalz, Edys

Problems in stabilizing the judgment process

In Quellmalz, Edys (Ed.), *Test Design: Studies in Writing Assessment*, Los Angeles, CA: UCLA, Center for the Study of Evaluation (1981),129-152 [ERIC Document Reproduction Service, ED 212 650].

Quellmalz worked for the Center for the Study of Evaluation, a research laboratory established in 1966 at UCLA. Here she takes up the "the renowned unreliability of judging constructed responses" (p. 1, the chapter's own pagination). She reviews "prevailing rating practices" and shows "how state-of-the-art rating processes pose serious threats to the validity of the writing assessments" (p. 2). Historically, her piece is the first full overview of issues of reliability in holistic scoring of essays. Since holistic scorers "rank essays by sorting them into piles anchored by the range of quality of that particular sample (Conlan, 1976)," the particular paper's rank would change with a different sample. "Such practices result in a 'sliding scale' where the rated quality of a particular paper changes according to the quality range of papers in the group" (p. 3). Anchor papers don't solve the problem because scoring procedures "still require raters to distribute papers across the score range" (p. 4). Quellmalz notes that anchor papers should be distributed in a new rating session to see how they compare, something researchers have not done (p. 5). She then brings up the issue of "rater drift" (pp. 7-8), caused by fatigue and regression to individual standards. She reviews possible solutions: use of third readers, pooling of independent scores, randomizing the order of essays to rate, and frequent checks on the rating. Another solution is to give examinations that are domain referenced. Rubrics could refer to "basic structural features of a discourse mode" (p. 11). She analyzes rater training procedures in Spooner-Smith (1978), Quellmalz & Capell (1979), and Baker & Quellmalz (1980), all showing rater drift, some of it toward more harsh scores as rating progressed. Conclusion: "Our rater drift comparisons suggest that total scores [e.g. summing of analytic trait scores] and a holistic score seem to mask fluctuations in judgments on the elements that contribute to the more global summary scores. We suspect that, at least during scale development and validation, assessments should collect separate ratings on component text features such as Support and Coherence that contribute to a total score. Otherwise, there is no way to identify and track consistency of the bases for global judgments" (p. 16). Quellmalz also notes a separate problem, the difficulty of holistic scoring to show writer improvement: "The holistic score provides no evidence of the developmental level of specific writing weaknesses that were low and may have improved" (p. 4). Quellmalz's survey of problems with holistic scoring was not bettered until Davida Charney (1984) and Catharine Keech, (1988, Part 1 and Part 2).

KEYWORDS: direct, assessment, essay-scoring, needs-analysis, critique, rater-reliability, ranking, sampling, anchor, research-method, rater drift, fatigue, pooled-rater, domain-referenced, rubric, leniency, gain, development

Applebee, Arthur N., Judith A. Langer, Ina V. S. Mullis

Writing: Trends across the decade, 1974–84. Report No. 15-W-01

Princeton, NJ: Educational Testing Service (1986)

Periodically, the National Assessment of Educational Progress (NAEP), government-funded, tries to capture the current state of school education across the US. It has proved a problematical endeavor, as those running it often admit. For instance, Rexford Brown was rater, researcher, and Director of Publications for NAEP from its first round of testing in 1969-1970. He developed a deep skepticism of holistic scoring of essays: "Holistic scorers need never explain what they are doing; and thus did holistic scoring achieve a certain amount of respect in our profession" ("What We Know Now and How We Could Know More about Writing Ability in America," 1978, p. 2). He helped NAEP develop and try primary-trait scoring with the second round in 1973-1974, as providing more information for educators. After about a decade of testing, Brown was hopeful, in part because of the information about writing that NAEP had been gathering, although that information was difficult to summarize in part because of the very changes in scoring over the years. A decade later, in 1986, Arthur N. Applebee, Judith A. Langer, and Ina V.S. Mullins took on the task in *Writing Trends Across the Decade, 1974-1984*. Extending their study back to the 1973-1974 assessment, the authors used directly comparable assessments to file a 10-year report. Since NAEP had not given up on holistic scoring of essays, their report analyzed what would become in the U.S. a distinctive combination of holistic and primary-trait scoring (pp. 67-69). Described as "Task Accomplishment," *primary trait scoring* was accomplished through scoring guides that isolated particular features of writing essential to task completion and then grouping those features into five levels of proficiency, from unrateable to elaborated. Described as a measure of "Overall Fluency," *adjusted-rater holistic scoring* was accomplished through chief reader and table leaders surveying the samples to be read and calibrating reader responses on a 6-point scale. No rubric was used, and judgment was determined by the general impression of a writing sample relative to others. In terms of validity, no specific construct model was provided. As the authors were careful to observe, "very little is actually understood about the impact of various writing assessment methods on achievement, and this relationship needs to be researched further before drawing any conclusions based on the NAEP data" (p. 59). In the absence of a construct model, the test designers focused on aligning task and scoring. The primary trait and holistic scores were therefore aligned with the tasks—the very embodiment of criterion-based assessment. For the primary-trait scoring, the rubric was based on the three tasks eliciting informative, persuasive, and imaginative writing; for holistic scoring, the sample papers used for reader calibration were selected from the papers at hand. Technically, the use of two types of scoring was brilliant in its desire to design an assessment that reflected writing as it was used at home, at school, and in the community. Reliability was determined by having 20 percent of the papers scored twice. Results were reported innovatively and transparently. The strategy of scoring had captured, as intended, related measures of overall fluency and task completion ability. (Note that the authors did not link data back to the first

NAEP assessment of 1969-1970. The reasons given were due to sample size—about 2,500 papers on one imaginative task scored by the primary-trait method and about 400 papers on a different task rated holistically. As they wrote, "Given these limited data and the fact that any subgroup trends from 1969-70 to 1974-79 would be based on only one imaginative writing task," So Applebee, Langer, and Mullins discussed trends only from 1974 to 1984 (p. 63). Evidence of fairness was demonstrated in terms of sub-group analysis by gender (male and female students), race/ethnicity (subgroups of black, white, and hispanic students), and region (Northeast, Southeast, Central, and West). The authors also considered the findings related to "educational progress." While at ages 13 and 17 all three sub-groups improved from 1979 to 1984, it was equally true that the percentage of students writing papers judged as adequate or better was substantially lower for the three subgroups on each task. Performance by gender and region were similar to student performance on the nation as a whole. Brown's sense of hopefulness had been correct. Led by members of the university writing community working in partnership with the educational measurement community, a meaningful report card for the U.S. in writing assessment had been created.

KEYWORDS: NAEP, criterion-based, fairness, group differences, holistic, interrater-reliability, primary-trait, Task design, school, data, trend, progress

The Handbooks

We now turn to the handbooks. While the sources of evidence vary, common among all these handbooks is that they are tales from the field. Beginning with William Boyd, a school teacher who had served on the Research Committee of the Educational Institute of Scotland in 1919, and ending with Edward M. While, who had directed the California State University and Colleges English Equivalency Examination in the US beginning in 1973, the authors of these handbooks are united by their singular emphasis on helping classroom teachers. In each, holistic scoring plays a central role in gathering evidence that, in turn, allowed inferences to be drawn about student writing proficiency and how to help students better it.

Boyd, William

Measuring devices in composition, spelling and arithmetic

London: Harrap (1924)

True to the UK tradition, no one has expressed devotion to teachers more clearly than William Boyd, a lecturer in education at the University of Glasgow. He recalled that the "main inspiration" of his work was "the thought of helping teachers in the daily work of the classroom by the improvement of teaching methods and by providing means for the more accurate estimation of the results of their work" (p. 7). In his section on measuring written composition, Boyd focused on the qualification examination in Scotland (similar to the 11-Plus examination in England and Wales). To support teachers, Boyd advocated impressionistic scoring to assess compositions. As he wrote, "Probably most teachers who mark essays … do not consciously concern themselves with the separate features marking up the whole" (58). Nevertheless, to help steady judgment he proposed "the essential qualities of a good essay (59) as divided into two

groups: mechanical (neat and legible script, correct spelling and punctuation, grammatical accuracy, fluency) and aesthetic (good vocabulary, good clause structure, good sentence structure, and effective arrangement of material). As to overall quality, separable traits may give clues. The surest sign of an essay above the average is unusual vocabulary, while the surest sign of an essay below average is failure to begin new sentences with a capital letter. But in the end, Boyd insisted on a gestalt-like sizing up of the whole essay. "An essay, like every other product of spiritual activity, is always more than the sum of its parts. We cannot ignore certain of the parts—spelling, for example—but if we are to mark justly as well as steadily we must always keep the whole in mind in our judgments" (p. 63). Thus did Boyd envision the construct of writing. To counter the alarming unreliability of teachers in marking essays (which Boyd demonstrated and found alarming), he created a 7-point *sample matching* scale, each point represented by a model essay. The essays were based on a general vote by readers. Teachers could then match a student essay with the closest essay on the scale, promoting "steadiness and promptitude" (86). The procedure is not holistic scoring by our definition, but Boyd did posit that the link between *pooled-rater scoring* leading to the use of a standard scale would increase interrater reliability. It is of interest to note that Boyd also found that "the companion inquiry in arithmetic revealed a similar uncertainty as to what is or is not satisfactory work in the subject" (p. 147). As with writing assessment, the problem was twofold: variation in texts and variation in marking.

KEYWORDS: analytic-holistic, pooled-rater scoring, gestalt, Scotland Qualifying Examination, interrater-reliability, school, Scotland, 11-Plus, data, sample matching scale

Vernon, Philip E. (Ed,)

Secondary school selection: A British Psychological Society inquiry

London: Methuen (1957)

In the post-war years, the UK educational setting would change dramatically. The Education Act of 1944 would be built on the premise of equality of opportunity for all children in Britain and Wales. Local authorities would oversee three different kinds of schools, providing instruction and training in three categories: grammar (viewed as a road to university and the professions), modern (a less rigorous curriculum), and technical (focusing on spatial and mechanical aptitude). Children would be streamed into one of these three categories on the basis of an examination at age 11 or 12. The examination became known as the 11-Plus, 11+, or Eleven-Plus, or simply Selection exam. By 1957, the publication date of Phillip E. Vernon's *Secondary School Selection*, all children who had reached the age of 11 (but not 12) on September 1 were required to take a three-part examination—categorized as Intelligence, English, and Arithmetic—in their own schools on a fixed day. Adrian Wooldridge (*Measuring the Mind: Education and Psychology in England 1850–1990*, Cambridge University Press, 1994), finds in Philip E. Vernon the beginning of the end of Galtonian orthodoxy. By the late 1950s in the UK, hereditarian theory was yielding to behavioral psychology, and researchers were becoming increasingly skeptical of the use of IQ tests in the 11-Plus examinations as a means of predicting student performance. In terms of validity, Vernon and his colleagues (Vernon's committee had twelve members, including Stephen Wiseman) thus turned to the significance of identifying a

more adequate criterion of grammar school achievement. The report concentrated on the predictive power of the examination, with subsequent performance three years later appearing the best target for evidence. The report also concentrated on the overall validity of the examination itself. As the authors noted, the test of intelligence was related to, yet distinct, from the English and arithmetic sections. Evidence would therefore be needed regarding relationships among these three conceptually-related constructs. Vernon attends to methos of marking essays, despite their unreliability, because essay writing added appreciably to the predictive validity of the whole exam. It is worth noting that Vernon's survey of "general impression marking" in the UK (pp. 120-121) was widely known by US evaluators (e.g., Braddock, 1963, Godshalk, Swineford, & Coffman, 1966, above). In terms of intended positive consequences, the report marshals the major studies that supported, with qualification, the process of post-selection streaming to encourage the advantages of keen competition and to facilitate rapid educational progress. Regarding unintended negative consequences of test use, the report was cautious about the newer limited-response examination forms: "[T]he notion of measuring ability in English by a test in which the child may be required only to underline words is particularly repugnant to many teachers" (p. 123). The report was alert to the impact of the entire process of selection on the mental health and personality development of children. Indeed, the majority of the 32 recommendations made in the final chapter of the book may be broadly interpreted as dealing with issues of fairness regarding the limits of the examinations and the need for multiple measures of student ability.

KEYWORDS: Education Act of 1944, 11-Plus, fairness, Intelligence Quotient, mental health, predictive, validity, reliability, Britain, school, 16-Plus, washback, coaching, objective-testing, essay-testing, pooled-rater, general-impression, border-zone, assessment, review-of-scholarship

Dressel, Paul L. (Ed.)

Evaluation in higher education

Boston, MA: Houghton Mifflin (1961)

In 1958, Paul Dressel, Director of Education Services at Michigan State University, had published *Evaluation in the Basic College at Michigan State University* (Harper and Brothers). The volume covered the university's general assessment efforts (managed by a Board of Examiners) that had begun in 1944 under a central administrative unit (the Basic College). In 1961, Dressel issued a second volume, *Evaluation on Higher Education*. There Dressel and his colleagues provided what could be considered the first institutional research handbook. Viewing evaluation as an "integrative element" in postsecondary education, Dressel stressed that evaluation was both a means and an end to improve quality of instruction (p. 24). Taking Michigan State College as a case study in institutional research, Dressel and his colleagues provided chapters on the nature and role of evaluation; specific evaluation problems in the social sciences, natural sciences, the humanities, and communication; the relationship between grades and examinations; and the role of institutional research in planning and policy development. An appendix provided a discussion of technical considerations in measurement. While limited-response forms of assessment are present throughout the volume, notable is attention to writing in the disciplines. The chapter called "Evaluation of Communication Skills" (pp. 192-226) was

written by Osmond E. Palmer, who was directing the Examinations Board for the Basic College at the university. Palmer reviews three formal methods of grading or rating student essays: sample matching, analytical, and ranking. Ranking, of course, is the underlying evaluative structure of holistic scoring. Palmers says that the method "assumes that one is reading for over-all effective communication (or possibly a single major thing)" (p. 213). He notes problems with the halo effect and the reduction of range of scores due to unevenness of papers: "the papers good on some factors may be poor on others" (p. 213). Palmer had worked with Paul Diederich and his system of holistic scoring at the University of Chicago (see Diederich, 1946, above), and at the time that Dressel's handbook appeared was influencing the shape of the Educational Testing Service's validation the College Board's English Composition Test (see Swineford & Godshalk, 1961, above), which eventually led to Godshalk, Swineford, and Coffman, 1966, above). The overall conceptual system described in Dressel's handbook is the most informed and best presented in the handbooks produced from 1924 to 1985. In many ways, the system supports what may accurately be understood as the first Writing in the Disciplines handbook in the US.

KEYWORDS: program assessment, WID, Michigan State University, Basic College, evaluation, measurement, Osmond E. Palmer, essay-scoring, analytic, ranking, sample-matching, holistic, University of Chicago, Diederich, CEEB

Diederich, Paul B.

Measuring growth in English

Urbana, IL: National Council of Teachers of English (1974)

Taken in light of his 1974 classic, Paul B. Diederich's 1966 guide to program assessment ("How to Measure Growth in Writing Ability," *English Journal*, 55 (4): 435–49) is remarkable for at least one reason. Diederich conceptualized validity in terms of reliability, not so much between readers as "the amount of agreement between two sets of independent measures of the same characteristic in the same students, taken at about the same time" (p. 104). His preference for analytic scoring and his disposition toward an expanded notion of reliability informed what would become the most widely read handbook on writing assessment in the US in the mid-1970s. Note that in both works Diederich avoided the term *holistic*, except to define it as "a term not used in this booklet" (p. 100). *Measuring Growth in English* is notable for its extended discussion of bias—a feature that is important and often missed in historical accounts. With ETS colleague Benjamin Rosner, Diederich had conducted an experiment in which teachers were asked to grade papers that had been explicitly marked as either from "honors" or "regular" students. Rosner had, however, deliberately altered the information so that the opposite group had actually written the papers. The result was that, even though the papers were from traditional students, the papers marked honors averaged almost one grade point higher that they should have earned based on alternative marking with correct identification. (Diederich's account of Rosner's experiment appears in a draft of *Measuring Growth in English*, but does not appear in the book; see Diederich, 1973, ETS archives.) With this investigation into anonymity of authorship, we find an early concern with bias in writing assessment as "the influence on grades of irrelevant consideration's such as liking or disliking the student, disagreement with his views, etc." (p. 99). For more on one of the more intellectually complex and often underestimated researchers in US

writing assessment, see Robert L. Hampel's *Paul Diederich and the Progressive American High School* (Charlotte, NC: Information Age, 2014).

KEYWORDS: rater-bias, factor-analysis, interrater-reliability, improvement, measurement, rating, teacher-reliability, evaluation, interrater-reliability, grading, high-school, predictive, analytical, general impression, scale, factorial, data, holistic, criterion-referenced, rater-training, Benjamin Rosner, anonymity

Cooper, Charles, and Lee Odell (Eds.)

Evaluating writing: Describing, measuring, judging

Urbana, IL: National Council of Teachers of English (1977)

In 1977, the publication of *Evaluating Writing* heralded the fact that the composition profession already had, in fact, a body of knowledge about writing assessment. Methods existed that had been used for a variety of purposes, from generating a portrait of student ability through the National Assessment of Educational Progress to identifying mature word choice through computer evaluation. Chapters by Charles R. Cooper and Richard Lloyd-Jones treated, respectively, holistic and primary trait scoring—and therefore covered forms of evidence related to construct, concurrent, and predictive validity. Chapters by Patrick J. Finn, Kellogg W. Hunt, and Lee Odell linked writing ability to processes of maturity and critical thinking, thus providing discussions of conceptually related constructs. Mary J. Beaven provided a chapter of goal setting and peer-review, thus calling attention to self-efficacy associated with the intrapersonal domain of composing and collaboration. In many ways, this 1977 edited collection signals a potential broadening of the field's understanding of the writing construct in ways that had not been established before. Topics of scoring were broadened by attention to the span of the writing construct and the varied ways that writing ability could be investigated, from computer-generated lists of word frequency to asking students for self-evaluation of their writing processes and the effectiveness of their products. Cooper's chapter, "Holistic Evaluation of Writing" (pp. 3-31) proved enormously influential, even though few shared his sweeping concept of "holistic evaluation" as "any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing" (4). That eclectic definition allowed him to call "holistic" procedures such as *sample matching*, *analytic scales*, and *primary-trait scoring*. To this day, the book itself is recognized as a classic of its time.

KEYWORDS: computer-analysis, peer-review, syntax, evaluation, measurement, holistic, primary-trait, computer, feature-analysis, computer, development, self-evaluation, peer-evaluation

Myers, Miles

A procedure for writing assessment and holistic scoring

Urbana, IL: National Council of Teachers of English (1980)

When we interviewed Miles Myers in 2014 shortly before his death, he recalled that some people laughed about "holistic scoring" being a California term. As he wrote to us, "My response more than once was, 'Wait till you see our Levitation system!'" For Myers, the usefulness for holistic scoring was associated with Eleanor Rosch's work on categorization. As Myers explained, Rosch has been influenced by Wittgenstein's concept of family resemblances as introduced in his *Philosophical Investigations* (1953). Rosch's research on categorization became Myers' inspiration for holistic scoring systems. Identification of prototypes, not elements, was the key to holistic scoring. But basically *A Procedure for Writing Assessment and Holistic Scoring* (1980) was not a philosophical monograph. It was a practical manual for teachers and administrators to run their own holistic evaluations—a project that emerged from early Bay Area Writing Project efforts, beginning in 1974, to spread holistic scoring in California schools. There are guidelines on topic selection, conducting a scoring session, and writing the report for stakeholders. The term "validity" does not appear in the book, and reliability is dealt with only in terms of interrater consistency. Whatever faults might be associated with the volume are overcome by its aim. As Myers told us, we should "remember that the 'procedure' book was intended as a proposal for teachers to try scoring sessions in their districts, and we did always require that every step had to be taken. We wanted districts to fund teachers engaging in a collective examination of student writing in a school or district. BAWP had learned that collective efforts could become an ethical commitment to both reliability and validity, calling into question how grading typically worked." For Myers, assessment had a positive impact as teachers bonded together to form communities dedicated to moving beyond reductionist ways of teaching and assessing writing (see White, 1985, below).

KEYWORDS: interrater-reliability, evaluation, rating, holistic, primary-trait, manual, Bay Area Writing Project, prototype, Eleanor Rosch, anchor, rater-training, criteria, scoring-guide, rubric, scale, uneven, 'splitter', writer-reliability, pedagogy

Davis, Barbara Gross, Michael Scriven, Susan Thomas

The evaluation of composition instruction

Pt. Reyes, CA: Edgepress (1981)

With a Doctor of Philosophy from Oxford University, Michael Scriven had held the rank of professor at the University of California, Berkeley when he conducted his evaluation of the Bay Area Writing Project (BAWP), from 1977 to 1978. According to Miles Myers in his 2014 interview with the authors, "BAWP needed Scriven, a professor of philosophy, almost as much as Scriven needed BAWP." An external evaluation of the program by an eminent evaluator would lend credibility to BAWP, a fledgling educational development program that needed external funding, and would give Scriven the kind of momentum that, in fact, earned him the rank of University Professor and Director of the Evaluation Institute at the University of San Francisco in 1978. In his e-mail to us, Myers quoted from Scriven's 1980 Executive Summary of the Bay Area Writing Project and his conclusion that the program "appears to be the best large-scale effort to improve composition instruction now in operation in the country and certainly the best on which substantial data are available." Following the report, Scriven recruited two other authors (Barbara Davis and Susan Thomas) and a nine-person review panel (including Paul

Diederich and Myers himself) for *The Evaluation of Composition Instruction*. With the publication of the program-evaluation handbook in 1981, it was clear that the work had paid off: Examples from the BAWP were the core of the book, and it shares the teacher-oriented philosophy of BAWP. Both holistic and analytic scoring of essays are recommended, but the problems of holistic scoring are underlined: "Holistic scoring does not measure or provide information about particular factors that might contribute to effective writing. Further, holistic scoring yields only limited information (a total point score) which is not very useful for formative purposes. In addition, the general qualities that composition teachers tend to weight in holistic scoring may be only remotely related to the commonsense requirements of functional prose writing, because teachers may be preoccupied with stylistic considerations" (pp. 89-90). Scriven, it should be noted, has supported the analytic or checklist approach to assessment all of his professional career. In fact, while evidence related to validity, reliability, and fairness are stressed throughout the book, its design is structured around a checklist approach based on phases of the evaluation: previewing, designing, conducting, synthesizing, reporting, and evaluating. In turn, these phases are informed by "scientific design" with attention to test construction and comparison groups (p. 29). It is, in fact, just this orientation toward design that flummoxed Myers. In his analysis of the BAWP, we can see the quintessential Scriven in action as he classifies gains in professional development as side effects—those "unintended good and bad" aspects of the instructional program that could be determined by brainstorming with program administrators, reviewing similar projects, analyzing the data at hand, observing instructional activities, and interviewing national leaders. Using these techniques, Scriven reports the side effects he observed in his program review of the BAWP: increased professional status of teachers, disciplinary integrity of writing instruction, pedagogical shifts to a process-based view of writing, increased collaboration among school districts and postsecondary institutions, improved monitoring of outcomes, and model use in other disciplines. Of course, to Myers, these were the very basis of the BAWP, which explains why the Scriven assessment model was quintessentially different from that needed to evaluate the project. (For a different approach to writing-program evaluation, see Witte & Faigley, 1983, below.)

KEYWORDS: evaluation, teacher-training, criteria, program-validation, program-evaluation, student-attitude, teacher-attitude, portfolio, administrating, program-cost, analytic-holistic, pedagogy, formative, summative, checklist, teacher training

Jacobs, Holly, Stephen A. Zinkgraf, Deanna R. Wormuth, V. Faye Hartfiel, Jane B. Hughey

Testing ESL composition: A practical approach

Rowley, MA: Newbury House (1981)

Content, organization, vocabulary, language use, and mechanics—these were the criteria or traits of writing that stood as the unifying framework for the three volumes written by Holly Jacobs and her colleagues: *Testing ESL Composition: A Practical Approach* (1981), *Teaching ESL Composition: Principles and Techniques* (1983), and *Learning ESL Composition: A Workbook* (1985). A young reviewer of this deeply theorized trilogy, Liz Hamp-Lyons, noted that the assessment volume had been influential in the US, but had, unfortunately, been ignored

internationally (*Language Testing* 1.2 (1984), 241-244). Believing that the first volume was a milestone achievement, Hamp-Lyons praised it as the first time that a review of assessment in both first and second language writing was provided between two covers, accompanied by details of the method as used at Texas A&M University, College Station, and by practical advice for implementation. The method, the ESL Composition Profile, is profoundly antithetical to holistic scoring. It preserved, for teacher and student, the features or traits on which a global score had been based. The breakdown followed a classical rhetorical framework of inventio (content), dispositio, (organization) and elocutio (vocabulary, language use, and mechanics). Each variable was categorized by four levels of proficiency, ranging from excellent to very poor. The profile consisted of scores for each criteria and the sum of scores for a total score. The separate components provided diagnostic information, while the total score provided an index of the writer's overall success at composing. Historically, the trilogy of volumes is widely recognized as influential in teaching and assessing second language writing. It clearly influenced Hamp-Lyons and her work, 1982-1984, instituting a profile score for the British Council's Proficiency Test of the English Language Testing Service (see Hamp-Lyons, 1987, 1992).

KEYWORDS: analytic-holistic, construct-validity, interrater-reliability, ESL, testing, measurement, assessment-procedure, scale, multiple measure, English Composition Profile, assessment profile, reliability, validity, data, Hamp-Lyons

Witte, Stephen P., Lester Faigley

Evaluating college writing programs

Urbana, IL: Conference on College Composition and Communication (1983)

Stephen P. Witte and Lester Faigley were aware of Michael Scriven's 1981 handbook, *The Evaluation of Composition Instruction*, co-authored with Barbara Gross Davis and Susan Thomas (see above). They were skeptical of Scriven's "goal-free model" that does not focus on program objectives for fear that they will bias the evaluator and lead analysis away from important "side effects" (p. 57). "Scriven's arguments in favor of goal-free evaluations notwithstanding," Witte and Faigley write, "most evaluators try to identify the goals and objectives of the program during the beginning stages of the evaluation" (p. 57). For them, there was no sense of scientific objectivity that could be achieved by bracketing the aims of the program and seeking information from comparison groups. Their deep sense of contextuality and contingency is evident in their statement about holistic scoring, that "judgments of writing quality are always relative. Raters give a particular score to a particular paper in relation to the scores assigned to the other papers in the set. A holistic training session might be defined as the process by which experienced raters of student writing are forced through group pressure to abandon their own ideas of writing quality and to adopt others which are relative to the rating group's view of writing quality, relative to the set of essays being rated, and relative to the need to distribute essays across all scoring categories. Whenever ratings—whether holistic or otherwise—are made relative to all the papers in the set, all rating are based on explicit and implicit comparisons among the papers in the set" (p. 15). Relativism is the order of the day. In light of this perspective in which context and change are acknowledged, Witte and Faigley encourage educational program evaluators to pursue a "paradigm of choices"—as Michael Quinn

Patton had termed the newly emerging call for multidisciplinary in the social sciences (*Qualitative Evaluation Methods*, Sage, 1980, p. 20). In Patton's recognition that "different methods are appropriate for different situations," Witte and Faigley found a way forward for their own work. Theirs was a component-based approach based on five elements: cultural and social context, institutional context, program structure and administration, content or curriculum, and instruction. Assessment of student writing might play a minor or ambiguous role in the comprehensive assessment of program. In their review of their own attempt to study improvement in writing at the University of Texas, they conclude, "After two years of collecting and analyzing over 28,000 scores on various measures and categories, we discovered that we had found out little or nothing about what instructional practices or what composing practices brought about the higher holistic scores at the end" (p. 34). While aspects of validity and reliability are discussed through the volume, Witte and Faigley place emphasis on a heuristic that derives from analysis of the five components and their interactions. The model offered by Witte and Faigley was an inspiration for the Design for Assessment (DFA) model for program assessment developed three decades later by Edward M. White, Norbert Elliot, and Irvin Peckham (*Very Like a Whale: The Assessment of Writing Programs*, Utah State University Press, 2015).

KEYWORDS: evaluator questions, program assessment, program-validation, program-evaluation guidelines, multiple measures, quantitative, research-method, contextual, change, holistic, contextuality, norm-referenced, rater-training, norming

White, Edward M.

Teaching and assessing writing
San Francisco, CA: Jossey-Bass (1985)

"It's about time." David Taylor wrote in his review of Edward M. White's *Teaching and Assessing Writing* (*College Teaching* 33.3, 1985, pp. 140–141). While a fifteen-year "flowering of research" may have caused a revolution in the ways that writing was taught and evaluated, that information had been "trapped in journals, and scattered across one-issue research monographs and anthologies" (p. 140). Praising the publication of single-author guide, Taylor noted how the book's theme—that sophistication in assessment improves writing instruction—was both innovative and promising. Especially noteworthy was White's emphasis on holistic scoring that could be connected with all aspects of classroom teaching—a marriage that White believed, according to Taylor, was heaven made. In this influential volume based on administrative and research experiences in California begun in 1973, White offered a well-written volume filled with research narratives useful for anyone seeking what was then a fresh view of how instruction can be informed by assessment. A scholar of Jane Austen who recognized the importance of narrative and irony, White knew how to weave a handbook. He covered fundamental principles of measurement involving validity, reliability, and fairness while elaborating on naturally occurring issues involving instruction and its relationship to assessment, Notable was his presentation of "controlled," *rater-adjusted holistic scoring* as a humanistic response to assessment. Alert to poststructural theories of reading that would appear, at first glance, to have no part in holistic interpretation, White saw these as a "resistance movement" against narrow analytic interpretations (p. 92). The process of reading, especially in

interpretative communities, was to be emphasized over what used to be considered the "true" meaning of a text. By the time readers reached the last sentence of the last chapter, they had been inductively led the very place White began: "The more we know, and the more we help our students know about assessing writing, the more effective our teaching will become" (p. 289).

KEYWORDS: assessment, evaluation, pedagogy, holistic, adjusted-rater, proficiency, improvement, assignment-design, essay-exam, reading-theory, reader-response, response, portfolio, measurement, program-validation, political, controlled holistic scoring, rater-training, improvement, handbook