CONTENTS

1

NINE MAPS

True maps are made of experiences.
—Paulette Jiles, *North Spirit*

In the United States, holistic scoring of student essays spread like a diet fad. During the mid-1960s some testing experts were still gingerly handling the upstart term *holistic* with scare quotes (e.g., Godshalk 1966, 2). Twenty years later, nearly every English teacher knew what the method entailed. Or thought they knew. One incontestable historical truth about holistic scoring is that from its beginnings, it was not monotypic. Despite persistent efforts to sway it toward that condition, the activity perversely remained polymorphous and indeed grew more and more so. Holistic scoring seemed to have been owned by the imp of the perverse.[1]

Practitioners, for instance, kept hunting and never finding a method of scoring essays that achieved high levels of interrater agreement leading to high interrater reliability coefficients. No matter how rudimentary the rubric, how long the training, or how strict the monitoring, the experience of doing holistic scoring varied according to writing task, individual scorers, and those scorers' singular responses to the scoring environment. One of our informants, who supervises online rating of a commercial essay test, told us that the morning after the US presidential election in November of 2016, the agreement of raters "went to hell," and that was using a simple four-point scale. As we show in chapter 8, when research began exploring the "ecology" of holistic scoring, a panoply of affects was shown to influence silently the agreement rates of individual scorers (Lucas 1988).

In fact, at every point in the art and act of holistic scoring, complexity reigned. This fact should not surprise. Whatever its stripe, holistic scoring is not one activity but a nexus of many. It is much more than just a technique for rating texts, a procedure in which people can be trained. As this book illustrates, holistic scoring acts as a ritual for crossing social and educational boundaries, an expression of symbolic power, a service

DOI: 10.7330/9781607329121.c001

for hire and a commodity for sale, a chance for in-group camaraderie and solidarity, a tactic to meet political demands for accountability, a shibboleth of insider taste and knowledge, a temptation to join a bandwagon or follow a craze, a tactic within a plan of research, an academic classroom practice, an act of psychological perception and interpretation. This list is random and truncated. The course of holistic scoring is entangled without recourse in its human enactments.

THE CONCEPT OF ORGANICISM

The history of holistic essay scoring, however, has one clear-cut moment, in 1926 when Jan Smuts published *Holism and Evolution*, thereby coining the terms "holism" and "holistic." The book by Smuts covered all of nature, physical and biological, and his concept of holism rested on a broad intellectual movement that extended back at least to European Romanticism. For simplicity's sake, let's call that movement *organicism* (although other terms would work as well—*holism, structuralism, field theory, systems theory*). The core energy of organicism opposed a Cartesian, part-focused explanation of things. Organicism responded with its own explanation that focused on wholes and their properties and functions. As Smut put it in his beguiling style, "The creative intensified Field of Nature, consisting of all physical organic and personal wholes in their close interactions and mutual influences, is itself of an organic or holistic character—that Field is the source of the grand Ecology of the Universe. It is the environment, the Society—vital, friendly, educative, creature of all wholes and all souls" (1926, 354–55).

In the decades when holistic essay scoring was starting up, the decades covered by this book, organicism flourished in all professional fields: anthropology, architecture, art studies, biology, education, environmental studies, learning theory, neurology, nursing, philosophy, psychology, psychiatry, religious studies, social work, sociology, the list goes on. Most appropriated the term *holistic* as soon as Smuts offered it to the public. Holistic essay scoring was a small plant growing in an intellectual field teeming with holism.

Watch organicism prosper, for instance, in intellectual domains close to holistic essay scoring. Psychology: in 1932 a review of "early holistic psychologists" notes that they had "much in common with the modern proponents of the configurationist viewpoint . . . a more organismic, phenomenalistic, and 'holistic' approach to the study of psychology" (Commins 1932, 217). Psychoanalysis: in 1943, Abraham Maslow is promoting his concept of the psychic "syndrome" as a whole that "can be seen in any of its parts if these parts are understood not reductively, but holistically" (528). Learning theory: in 1954, Louis Thorpe and Allen Schmuller include in their textbook of learning theories a chapter called "Gestalt Psychology: A Holistic Outlook," which offers the fundamental principle "that the organism reacts to total situations and proceeds from whole to part and from general to specific on the basis that the whole is always greater than the sum of its parts" (247). Linguistics: by 1963, in a collection called *Parts and Wholes*, linguist Roman Jakobson is warning about the danger of the *pars pro toto* fallacy in analysis of language, "the illicit conversation of a mere part into a seemingly self-sufficient whole," resulting in "the artificial treatment of messages without reference to the superposed context" (159). Education: by 1967, educationist Paul Hanna is defending his theory of spelling on the grounds that it includes phonemic, graphemic, morphological, and syntactic cues, in short, assumes a "holistic language structure" (216). Literature: by 1974, literary critic James Bennett is editing a special issue of *Style* called "Holistic Criticism," "a conceptual framework for the full description of the dynamic design of a literary text (syntagmatic and paradigmatic) as it connects with author, audience, and world. And, I should add, with the critic" (288). Composition and rhetoric: by 1975, Joseph Comprone is arguing that first-year college students should study "holistic" communication systems and that the study of rhetoric should "adopt the language of quantum physics, gestalt psychology, dialectics, cybernetics, general system theory, or half a dozen other disciplines. One could describe a shift of emphasis from stasis to process, entity to relationship, atom to gestalt, scaler to tensor, component to system, analytics to dialectics, causality to constraint, bioenergetics to communication, or at least a dozen other parallel shifts" (2).

These fields have varied degrees of closeness to writing assessment, but that proximity is another book for others to write. Except for the next chapter with its history of gestalt perception theory, we provide no systematic chart of the vast country of organicism. Even in our restricted terrain of holistic essay scoring, however, history makers and history readers can get lost. It seems a map is needed, but one map won't do. We need a portfolio of maps. Here are nine.

MAP 1: CONSTANCIES IN THE HISTORY OF HOLISTIC ESSAY SCORING

Over the decades this book covers, some constancies do emerge. One is the fact that today in both the United Kingdom and the United States, the universal popularity of holistic scoring has markedly diminished, presenting a pattern of rise and decline historians and their readers find familiar.

Running opposed to this decline and fall, however, is a second constancy, the uninterrupted ascent in national attention to literacy proficiency.[2] Educational accountability almost always includes measurement of writing proficiency. After WWII there was a striking growth in the practice, study, and discussion of formal assessment of writing. In 1949 no professional journal devoted itself solely to educational evaluation or assessment; by 1999, eight did.[3] Over the same span of time, the volume of published scholarship on evaluation of writing increased thirtyfold. The newcomer on the block, holistic scoring, received especially intense scrutiny on both sides of the Atlantic. Did the scrutiny eventually work against the establishment of the method?

The question, answered in chapter 9, leads to a third historical constancy, which was the preoccupation of scholars and practitioners with one particular aspect of the rating of proficiency or performance in writing: the formal methods by which that rating can be accomplished. So those methods also grew, at least in number. And as they grew, they competed. What works better—counting misspellings, scoring multiple-choice questions on grammar and vocabulary, appraising the revision of a poorly written text, turning the student essay into a cloze test, applying a content scheme, or summing up the scaled ratings of eight different writing traits? In formal holistic essay scoring, two or more independent raters label the worth of a whole essay with a single score and do so without first assigning numerical value to separate accomplishments of the writing (recall the definition provided in the introduction). The mystery is why, over many decades, holistic scoring connected so centrally to every other method of writing evaluation—connected in theory, practice, ethics, cost, doability, and faculty esteem. How did that centrality come about? And what happened to it? And is it gone forever?

MAP 2: HISTORICAL EMPLOTMENTS

Simply put, this book studies the early histories of holistic scoring of writing in the United Kingdom and the United States from the mid-1930s to the mid-1980s. Histories, in the plural.

This book disavows a single unified history of holistic scoring, however complex that one history might be depicted. Holistic scoring has many histories because history itself is constructed. The human past was made by humans, true, but history is made by historians. In a phrase current around the same time our own account of US holistic scoring starts, history is an "imaginative reconstruction" (Dray 1963, 108–10).[4] The English philosopher R. G. Collingwood captured the gist of this concept in one sentence: "Historical thinking is that activity of the imagination by which we endeavor to provide an innate idea with detailed content" (1946, 247). The nub of this definition is the word *innate*.

One innate or a priori idea historians bring to their history making is a sense of time passing. This fact is not entirely self-evident. Historians may suggest that temporality inheres in the past, but actually it inheres in the telling of the past. In his summary of constructivist historiography after WWII, Donald E. Polkinghorne puts this point deftly: "Historians have to set forth their presentations as sequences of events, which gives the impression that their conclusions are inferences from the evidence, when really they are only indicators of the way the evidence has been ordered" (1988, 52). Note Polkinghorne's shift from "sequence" to "order." History is not one thing after another nor even one thing over and over. It is one thing seemingly connected to another.

History is a sequence, cohering through narrativity, plot, or in Hayden White's eye-catching word, "emplotment." And, as White famously argued, emplotment comes in different configurations. The historian can narrate the French Revolution as a triumph over repressive rule (romance), a flowering of a union between youthful natural energy and aged social order (comedy), a failed effort to achieve a better society (tragedy), or a hopeless struggle against inevitable forces (satire) (1973, 45–80).[5] Nested within White's "archetypes" or "governing metaphors," and more directly shaping the accounts of most historians, can be read scores of more specific narratives, storylines usually widespread in the historian's own day and culture, plots such as rags to riches or pure intentions corrupted by material realities. The personal anecdote, seemingly unique ("Once I . . ."), usually takes shape in the form one or more of these cultural storylines.

The point is worth dwelling on since it is crucial to this book's particular method of seeking and displaying its history. We glimpse five loose and intermeshing narrative structurings.

- **History as annals**. To begin, we have tried to gather new data connected with holistic scoring in its early days. We have scoured archives, interviewed survivors, tried to sort the accurate from the inaccurate, recorded names, places, and dates—all to construct a new annals—a day-by-day record of an important period in UK and US education.
- **History as story**. But of course, we could never escape from narrative. If, as Collingwood and Pilkinghorne say, historicizing begins

with innate orderings, the search itself for new data will be constructed. Fingering through a dusty filing cabinet is usually searching *for*. We found that with holistic scoring, stories abound, both anecdotes and cultural plotlines. But our finding was part of emplotments we ourselves were writing. Who can resist a good story?

- **History as trajectory**. We mean by historical trajectory a segment or unresolved span of an emplotment. Naturally, the shorter the historical narrative stretch, the greater the chance it will be unfinished. Indeed, it can be argued that historical storytelling differs in one essential way from fictional storytelling in that history is never resolved. Unsettling as it might be, history is nonteleological. In her 1999 end-of-the-century account of US writing assessment, Kathleen Blake Yancey plots a hopeful trajectory of three progressive waves, from objective testing to holistic scoring to portfolio scoring, but who can say that difficulties with innovative assessments won't bring back multiple-choice testing?[6]
- **History as social exploration**. As was the case with work preceding this volume (Elliot 2005), the present volume assembles circumstances surrounding key assessment episodes in order to place them in their social context (316). Because the early practice of holistic scoring had important stakeholder consequences, its social history and social justice are deeply related. We realize that how we write our histories matters, and we encourage new interpretations of our social explorations that focus on impact. Especially promising here is the social justice historiography proposed by J. W. Hammond (2018), with special attention to stakeholder representation, measurement consequences, critical reflection, and practice implications. While this volume attends to each of these analytic frames, extended social justice historiography allows extended interpretation of the history we recount and the creation of assessment histories yet unwritten.
- **History as global journey**. In her study of contemporary trends in historiography, Eileen Ka-May Cheng (2012) finds that the most important recent methodological development is the shift to a global perspective. Along with an emphasis on narrative and cultural context, recent scrutiny of world connections has allowed historians to focus on specific events, such as those scoring genres described in our book, while locating them in a transnational context. We return to the need for a global perspective based on demographic trends in chapter 10. Meanwhile, we hope readers will welcome the transnational stories we tell and the revisionist potential they hold for histories of the profession and, by extension, for fair opportunities for all within the profession.

The danger with historiographic emplotments is that they may exclude some historical evidence and, worse, may curb the search for more evidence. "Every story that is told obscures the stories that go untold," writes Verlyn Klinkenborg (1992, 5). We add that every story told tends to obscure other stories that, often contradictory, lie hidden within it, waiting to be told.

MAP 3: FOUR VIGNETTES

These points beg for illustration. Four vignettes of holistic scoring follow, picked and condensed from the many we reconstruct in this book. Think of them as movie trailers.

Vignette 1: 1940, spring, England. With little warning, the school population of County Devon in the west of England suddenly doubled within a few months. The new students were largely evacuee children from London, sent to the country for safekeeping from the predicted German blitz. At the time, R. K. Robertson was Chief Examiner for Devon, in charge of the 11-plus examinations, mandated assessment that largely determined whether children eleven or twelve years old would complete their formal schooling (Wiseman 1929, 205n2). For an examiner faced with a huge increase in exam-taker population, Robertson headed in a most unexpected direction. In the language-proficiency part of the exam, all students took an objective short-answer examination. Then the roughly 2,500 of them who had scored in the middle third were given a second examination. It required them to write an impromptu essay that was then rated on a thirteen-point scale by four independent teachers. The raters read very quickly, and their rates were averaged. Robertson's method of scoring essays had been investigated for decades in England but had never been used in large-scale governmental examinations. Called general-impression marking at the time, it was the scoring method we term pooled-rater, and we see no reason it doesn't deserve the term holistic.

What is the narrative here for a history of essay-scoring methods? The local story finds home easily enough, a necessity-is-the-mother-of-invention tale or perhaps a good-comes-of-war story. An ironic version might be found in the story that in 1941, Pearl Harbor ended the essay the College Board had used for their admission test since 1900 (Elliot 2005, 99–101). The narrative, however, also fits into a longer trajectory. Robertson's method of scoring was borrowed by US testing experts in the 1950s and 1960s. So the plot is of emigration and taking root in a foreign land. This narrative looks quite different from the plot currently favored by US composition scholarship, that holistic scoring was homegrown, developed by the US testing industry by Educational Testing Service (ETS) employees Fred I. Godshalk, Frances Swineford, and William E. Coffman and published by the College Entrance

Examination Board in 1966.[7] In fact, for that experiment Godshalk, Swineford, and Coffman used a pooled-rater scoring method very similar to Robertson's. We return to this story in chapter 3.

Vignette 2: 1966, Princeton, New Jersey, and the University of Connecticut. The same year the College Board published Godshalk, Swineford, and Coffman's ETS study, ETS was sponsoring an Invitational Conference on Testing Problems. One of the presenters was Ellis Batten Page, who gave an interim report on his and Dieter Paulus's experiment at the University of Connecticut in scoring student essays by computer—a venture first funded by the College Board. To secure a trustworthy predictor variable, Page and Paulus had human raters score each of the essays to be analyzed by their FORTRAN program. They used an analytical approach, scoring five criteria separately on five-point scales and summing the scores (Page and Paulis 1968). This particular five-criteria scale, one that became enormously influential in the following decades, had been developed by an ETS researcher, Paul B. Diederich, and was first published in the *English Journal*, also in 1966. So the same year marks the public announcement of three distinct scoring methods: Diederich's analytic scale, ETS pooled-rater holistic scoring, and computer analysis of written communication.

The evidence undermines the common cultural story that within organizations—such as ETS and big science university programs of research—innovations arise through coordinated teamwork: a win-the-prize-by-rowing-in-unison story. Just as disorienting is the trajectory the story pieces out. It doesn't look like a one-thing-after-another history. It doesn't much fit progressive emplotments Hayden White called "romance," histories that narrate one force being replacing by a better force, a storyline signaled by words such as *post* (posthuman) or *beyond* (beyond outcomes), or by metaphors such as "generation" (Guba and Lincoln 1990), "turn" (Trimbur 1994), "phase" (White 2005), or "wave" (Behm and Miller 2012). Instead, the historical evidence suggests multiple and simultaneous plots, parallel and competing and unresolved, as in the first act of a play that cannot yet be taken as tragic, comic, or satiric.

Vignette 3: 1979, fall, University of Southern California. Louise Wetherbee Phelps joined the University of Southern California's department of rhetoric, linguistics, and literature and heard that faculty and graduate students were reading Subject A examinations, essays written by entering students for placement within or exemption from USC's first-year writing program. She asked to participate in their holistic scoring of the essays. But she lasted only one session. She just "couldn't obey their rubrics." She felt she "didn't belong" (pers. comm., January 7, 2015).

This personal narrative vibrates, antithetical, within the orthodox professional narrative that holistic scoring sessions, when run by faculty, give rise to a greater sense of camaraderie and solidarity. For instance, Carol Holder was coordinator for composition programs in the California State University system and involved with the scoring of its English Placement Test, and she remembers that "because of the fun faculty from different campuses had at the 3–4 day holistic scoring sessions, it wasn't hard to recruit scorers. We enjoyed wonderful dinners wherever we met—usually the San Francisco Bay Area—and developed friendships with faculty on sister campuses" (email to authors, March 4, 2015). But Phelps's experience is not really an anomaly, and this book records a number of similar personal anecdotes, some from scholars in the profession as eminent as Phelps became. Charlotte Linde, scholar of institutional storytelling, calls these stories "unspeakables": "Institutions have occasions that permit the telling of certain narratives. Other potential narratives, the unspeakables, are often difficult to speak because there is no sanctioned public occasion for them" (1997, 287). The underground resistance to holistic scoring forms a history of its own, as we recount in chapter 7.

Vignette 4: 1981, July, New Orleans. At a conference sponsored by the National Institute of Education called "Feasibility of Assessing Writing Using Multiple Techniques," discussions among school and college administrators revealed much confusion over terminology and scoring methods. One administrator heard descriptions of ETS holistic scoring and said that "what he used and called holistic scoring was nothing like the procedure developed by ETS." Other administrators agreed and described their own "unique" systems with pride. To them, other methods apparently had little appeal (McCready and Melton 1981, 80–81).

Is this evidence from a narrative of holistic scoring as fad? When narrating movements as popular as holistic scoring, it is easy to homogenize. Sociological research into popular trends, however, always finds a certain amount of internal conflict to explain their rise and decline. The peak of a fad may contain features that will lead to its decline: oversaturation, self-consciousness, misinformation, exaggeration, and departure from the original (Meyerson and Katz 1957; Miller 2013, 206–8). To this list, as noted above, we add resistance. Are these stages in the natural history of popular trends duplicated in some of the trajectories of holistic scoring 1949–1983?

We hope our reliance on narrative does not imply that this book will take lightly the profession's past agency connected with holistic scoring or will assume blithely that the current profession can only fictionalize those connections. With every sentence, this book tries to get as near to the actual territory as it can. Maybe history can draw only maps of the past, but maps can be more or less accurate, can lead people to where they want to go more or less well. Novelist and poet Paulette Jiles, who spent some years living in the "trackless" northern provinces of Canada, warns, "You must understand how useless a map is. You must also study them with great care" (1995, 192). In our analysis of splitters, rubrics, and profiles, we return to this theme in chapter 9.

MAP 4: EVIDENCE

Should historytelling be descriptive, explanatory, or predictive? Should it be qualitative or quantitative? Should it use a mixed or multiple methodology? Our answer to all these questions is "yes." To that methodological eclecticism, we added the ballast of old-fashioned documentary evidence. We have gone to newspapers, federal and state regulations, funded research reports, individual research articles, institutional and commercial testing archives, and current testimonials of those who were there.

- **Archives**. We located important and unanalyzed materials in archives, for instance, at the University of Chicago, Michigan State University, and ETS. The photo of James Britton in chapter 4 at the Dartmouth Conference, found for us in the Rauner Special Collections Library, may well be the only image that exists of that esteemed UK researcher at that event. Left unknown and untapped by us lies a vast body of evidence in archives around the world.
- **Structured interviews**. As we note in the acknowledgments, we interviewed people who were there—some face to face, some on Skype, some by email, some by telephone. We asked questions tailored to their individual experience. Taken from chapter 6, here is a question from our interview with Sydell T. Carlton on July 2, 2014:

  > One of the milestone events in the history of writing assessment is your famous study with John W. French and Paul B. Diederich on factor analysis in 1961. Can you provide the context for that study? Were its findings understood at ETS to be justification of a type of scoring that would focus on total reader impression—that is, a justification of holistic scoring?

  As might be imagined, the answers to such pointed questions helped paint the holistic enterprise in new, complicating, and contradictory detail.
- **The authors' experience**. There is no objectivity in history because history does not make itself. Humans make history by writing it. But historians can be closer to or further from the events, with advantages either way. The authors of this book were tangentially involved with holistic assessment only in the last years this account covers,

but they have not exactly maintained a hands-clean or heads-clear distance since. Although Richard Haswell was using holistic scoring for personal research as early as 1981, it was ten years later that he started considering and then modifying the method as part of a first-year and third-year campus-wide assessment of writing at Washington State University. Norbert Elliot worked for ETS during the summer of 1984, but it was a year later that he started applying and evaluating holistic scoring as part of a campus writing assessment at East Texas State University (now Texas A&M Commerce), 1985–1988. Throughout their careers, both taught, applied, and researched writing assessment. For good or bad, that experience is part of the baggage they brought with them to write this book.

- **Published empirical research**. Formal assessment of writing walks hand and hand with empirical evidence. By its nature, assessment collects and produces data. It shapes, records, validates, defends, critiques, and revamps itself through data. We took extra care to read the data-infused record, at first to explore how it, too, practiced, legitimized, and self-policed its culture. Slowly we became aware of an argument embedded in the data, an evidence-based perspective we present in the final chapter. Over the decades, in investigating and reporting holistic scoring, researchers enlarged their sources of evidence. The new evidence helped unpin holistic scoring yet perhaps at the same time pointed a way to revise and better holistic scoring. The empirical record, we argue, can help turn the history of a scoring method—however incomplete, in fact because it is incomplete—into a history that is actionable. History cannot be recounted just by re-counting it. We report numbers we were lucky enough to find, but we also sought out the underlying habitus and motivations of the people who produced the numbers. It is easy to see holistic scoring as shaped by numerical data—so many essays rated in so many hours with such and such a rater reliability. But holistic scoring was also shaped by philosophy, politics, economics, psychology, pedagogy, and personalities. In a phrase, this history pursues the praxis of the holistic, a praxis reproducing and validating the society and economy that privileged and sustained it. As students, teachers, researchers, and entrepreneurs were practicing the holistic, they were, chiefly unaware, practicing and legitimizing their culture (Bourdieu and Passeron 1990; Douglas 1986).

MAP 5: TERMINOLOGY

While examination and use of an evidence-based approach is the last step of this book, the first step was to take care of the terminology. From its start, discussion surrounding holistic scoring was plagued with synonymy and polysemy. A single concept or method attracted different names, and a name sometimes referred to quite different concepts or methods. This semiotic slide has been a problem since the early days of educational measurement when Truman Lee Kelley referred to definitional challenges as the "jangle fallacy" (1927, 64). Dwell a moment on any term provided in our glossary, and the next stop is a rabbit hole.

Everyone agreed, for instance, that *holistic evaluation* or *holistic assessment* dealt with the value of an essay taken as a whole rather than broken down by parts. But what does one call that wholeness? Around the end of the nineteenth century, *general merit* was popular, but later so was *total merit*, *general impression*, and *overall quality*. Philosophers, psychologists, linguistics, and educationists were adding *holism*, *wholism*, *interconnectedness*, *syndrome*, *template*, *agglutination*, and *glosso-dynamic* (Titone 1973). The adjective *holistic* could appear as *organic*, *organismic*, *nonanalytic*, *global*, *integrated*, *total*, *relational*, *configurational*, *unstructured*, *systemic*, *unitive*, and *molar* (Tolman 1932, from the Latin *moles*, a "whole"). The most common opposite terms were *analysis* and *analytical*, but synonyms spread like nonnative weeds: *atomistic*, *decompositional*, *featural*, *structured*, *synthetic*, *dimensional*, *trait-based*, *registered* (Braungart-Bloom 1984), and *meristic* (Bhatia 1977, from the Greek *merismos*, a "division").

Synonyms pose problems, but they are lesser problems. Far worse is when contemporary accounts use the same name to refer to different events, such as when *holistic* refers to scoring methods that, in fact, are analytical at root.[8] For our meanings of the host of technical terms writers are obliged to use when seriously discussing evaluation of writing, in application or research, we provide a glossary (and readers provide the rabbit hole). During the period of our study, we found some terms are often used and so are crucial to this book's map of the territory. The citations provided are typical uses in the United Kingdom and the United States. Let's begin with *analytic scoring* and move, by contrast, to *holistic*.

In the practice of analytic scoring, readers assign separate scores to different aspects or accomplishments of the writing. The scorers often use a checklist identifying the parts to be scored with scales for each part (Wiseman 1949). In using scaled criteria, readers fill out a scoring grid—in the United Kingdom, a *marking scheme* or *schedule*—that provides values for individual writing traits such as ideas, organization, or support, each on a scale from low to high (e.g., 1 to 4, or 1 to 20) (Diedereich 1966). Primary-trait scoring is also a scaled-criteria method in which the criteria are limited to a few relevant rhetorical requirements established by the writing task (Mullis 1976). For scoring short essays in academic-subject examinations, a content scheme—in Britain *mark scheme* or *marking scheme*—stipulates the points awarded for each relevant claim of the writer (Mather, France, and Sare 1965).

In the practice of holistic scoring, a scale is used to assign a single value mark to a whole essay—not separately to individual aspects. Holistic scoring is informal if there is only one rater per essay and if there are no preset schemes establishing parameters, such as anchor essays or a given distribution of rates (Gray and Ruth 1982). Typical informal methods are the time-honored grading of teachers; open ranking, in which essays in a set are simply rank ordered from worst to best; and sample matching, in which an essay is assigned a score or rank by fitting it within a given set of essays arranged from best to worst.

In distinction, formal holistic scoring—the subject of this book—compares the scores of two or more raters using a scale on each essay to provide a single value mark (Britton, Marten, and Rosen 1966). Formal holistic scoring ranges from open to controlled, depending on how well the scale levels are predefined by definition, description, or sample essays—"anchors," "range-finders," "exemplars,"—or, as Miles Myers (1980) put it, "prototypes." Two common but crucially distinct ways of controlling essay ratings are rubrics and scoring guides. The first arranges the selected criteria and numbered scale in a table format so each criterion is described and scaled in the same way (see fig. 9.2). The second describes the selected criteria in a way that connects to the holistic scale, but the connections are not uniform across the criteria (see fig. 5.2). Formal holistic scoring schemes can be imagined as closer or nearer to analytic schemes (see table 1.1).

Finally, in formal holistic scoring there have been three distinct ways the final score of an essay is calculated from the scores, often unalike, of independent raters. In pooled-rater scoring—also called "consensus scoring," "collective judgment" (Boyd 1924), or "multiple marking" (Head 1966)—scores based on a shared scale are simply summed or averaged, however many independent raters there are for each essay. Each score represents the perspective of one rater and is taken as no better or worse than another rater's score. In adjusted scoring each essay has two independent raters, but if their scores differ by more than a specified degree, a third rater is used to adjudicate and determine the final score of the essay, often during the reading itself (Breland and Gaynor 1979). In consulted scoring the two original raters discuss and resolve discrepancies (Pilkington 1967). Finally, in office-adjusted scoring a post hoc adjustment is made to reduce rater error (Hartog, Rhodes, and Burt 1936).

Returning to the jangle fallacy, we must emphasize the confusion that ensues when the major elements, defined above, are misnamed or confounded. At base, primary-trait scoring is analytic, and when it is designated "holistic" (e.g., Lloyd-Jones 1977, 37), it skews a narrative of holistic scoring that respects its perceptual underpinnings described in chapter 2. Again, by our definition, formal holistic scoring does not apply when student essays are read by only one rater with a supervisor occasionally spot-checking the scores, as has been standard with governmental school examinations in the United Kingdom (Office of Qualifications and Examinations Regulation 2014) and is growing more common in the United States as automated scores are used to provide a second score (Bridgeman 2013). We acknowledge the contested nature of our definition yet want our definitions to be clear lest our own interpretations become part of the jangle. With adjusted scoring, it is common practice to treat scores one point apart (e.g., 3 and 4 on a six-point scale) as a match and to treat as discrepant scores two or more points apart (e.g., 3 and 5). But when the scores of 3 and 4 are averaged within a group or scores of 3 and 5 are decided by a wiser third party, it is unwise to classify this mixture of pooled scoring with adjusted scoring under a single category of rater precision. Such classifications confound two different underlying philosophies of evaluation. Similarly, it is unwise to treat interrater reliability as a preferred, or (worse yet) only, measure of reliability.

As our book illustrates, many in the United Kingdom and the United States implicitly assumed they had provided evidence of validity by measuring degree of interrater agreement and interrater reliability. Evidence of validity, however, cannot be established by reliability measures alone, and those reliability measures themselves differ widely. Today, we think of reliability as characteristics of Differential Reader Functioning over Time (DRIFT): differential severity, differential accuracy, and differential scale category use (Wolfe et al. 2007). In early holistic reading reliability research, when a sample is only one twenty-minute impromptu essay from each student, early researchers radically simplify the notion of reliability, tacitly excluding basic questions such as the following: Would those same raters score the writing sample consistently a few weeks later (intrarater reliability)? If two tasks are given, are the forms parallel (test reliability)? Does the writer perform consistently on different tasks—and, if so, how many tasks are needed to make a claim about writing proficiency (writer reliability)? Each of these questions has implications for both validity and fairness.

We leave the glossary to provide definitions of other elements related to formal holistic scoring as they were used during the period of our study. As these terms are encountered, it is important to remember that the specialized vocabulary used in this book follows standard usage of the period from the mid-1930s to the mid-1980s in the United Kingdom and the United States. This historical resection means definitions do not include significant change in terms beginning in the late 1980s, as is the case with evidence related to validity (Messick 1989). Only in chapter 10 do we shift to contemporary definitions in order to propose an actionable future for writing assessment. Let chapters 2 through 9 therefore serve as evidence of the history-boggling permutations and combinations of these technical terms as they were applied by actual first-generation researchers working with methods of holistic scoring.

MAP 6: A DEFINITION

These terms provide context for our definition of formal holistic scoring of essays, whose early history this book will construct.[9] We define formal holistic scoring as *the use of a scale to assign a single value mark to a whole essay and not separately to individual aspects, with scorers trying to apply the scale consistently, and with the final score for each essay derived from two or more independent ratings*. Ancillary to this definition is the number of scale levels, identification of a scale as interval, degree of openness or control of rater training, use of rubrics or scoring guides or anchor essays, number of independent scores per essay beyond two, and scoring calculation involving pooling, adjusting, and consultation methods.

MAP 7: A CONTINUUM OF POPULAR ESSAY-SCORING PROCEDURES

Over the decades, several holistic scoring procedures stand out. Table 1.1 arranges nine of the most popular along a continuum from holistic to analytic.

Four observations are important in terms of table 1.1. First, often these methods are distinguished by number of raters per essay, a matter of historical importance. For instance, the United Kingdom did not adopt a two-rater method during the history we present in our book—and does not universally support one today. Following Cyril Weir's own analysis of the British history of interrater reliability, Weir, Ivana Vidakovic, and Evelina D. Galaczi note that, despite "a growing consensus in the profession on the need for and value of double marking . . . practicality is still proffered as an excuse for not utilizing this means of improving scoring validity even in the 21st century" (2013, 201). As a result of a program of research beginning in 2008, researchers and policymakers in the Office of Qualifications and Examinations

Table 1.1. Popular essay-rating methods arranged along a continuum of holistic to analytic

Holistic, open pooled-rater	Raters score each essay relying solely on a shared scale, with all scores summed or averaged (e.g., Britton 1963).
Holistic, controlled pooled-rater	Raters score each essay using a common scale and guidelines such as anchor essays, scoring guides, rubrics, or a given distribution of scores (e.g., Godshalk, Swineford, and Coffman 1966).
Holistic, office adjusted	Raters score each essay with scores later adjusted by standardization techniques, such as applying the overall standard deviation to each score (e.g., Hartog, Rhodes, and Burt 1936).
Holistic, controlled adjusted-rater, with anchor essays	Raters score each essay independently with a third reader resolving discrepant scores and a sample essay to illustrate each scale level (e.g., Myers 1980).
Holistic, controlled adjusted-rater, with scoring guide	Raters score each essay independently with a third reader resolving discrepant scores, using a guide that roughly describes each scale level (e.g., White 1973).
Holistic, controlled adjusted-rater, with unscored rubric	Raters score each essay independently with a third reader resolving discrepant scores, using a tabled checklist of specific writing traits but without points assigned for each trait at each scale level (e.g., Bossone 1969).
Analytic, primary trait	Raters score an essay on a restricted number of traits that are appropriate to one rhetorical requirement of the essay topic, with each trait given points on its own scale (e.g., Mullis 1976).
Analytic, scaled criteria	Raters score an essay using a tabled checklist of specific writing traits that have points assigned for each trait at each scale level, and the final score is the sum of all those points (e.g., Diederich 1966).
Analytic, profile	Raters score an essay using a tabled checklist of specific writing traits that have points assigned for each trait at each scale level, and the final score is the sum of those points (e.g., Hamp-Lyons 1987).

Regulation (Ofqual), a nongovernment regulatory department in the United Kingdom, concluded in 2014 that there is “a strong body of evidence from the 1940s to 1980s that double marking is a more reliable method of marking than single marking.” Nevertheless, Ofqual (2014) notes the “significant logistical and financial challenges associated with the implementation of double marking” (10). As a result, “none of the exam boards offering general qualifications in England currently use double marking in its true sense. Instead, all choose to quality assure marking through a sampling approach” (6). In this approach a student essay is scored only once, by a junior marker, whose scores are occasionally checked (“sampled”) by a senior marker. In the United States, however, multiple marking in formal assessment was legitimated early and largely remains. Examination of cultural values and subsequent historical interpretation are thus related to the number of raters in a given writing-assessment episode.

Second—and here is an important similarity between the two nations—the essay-rating schemes are directly related to evidence of validity. As Hartog, Rhodes, and Burt argued, "No test can be a 'valid test' unless it yields consistent results in the hands of different examiners, i.e., unless its 'reliability' (to use the word generally employed by educational psychologists), or, as we should prefer to say, its 'consistency,' is 'high'" (1936, 68). It is not simply that reliability is a prerequisite to validity in our studies; rather, it is that evidence of reliability stands as evidence of validity. The forms of validity that were used from the mid-1930s to the mid-1980s on both sides of the Atlantic—which we inferred from construct, content, concurrent, predictive, and conceptually related evidence—often emerge as a concern, albeit too often tacit and too often related to reliability, of the researchers we examine. In similar fashion, fairness is based largely on evidence of scorer reliability. Research into factors of gender and ethnicity, for instance, began quite late in the 1970s (e.g., Breland 1977; with the exception of Martin 1972). Again, we return to Hartog, Rhodes, and Burt in their belief, present for most of our history, that a "fair decision" involves the elimination of random variation related to scoring reliability (1936, 235). Therefore, while our understanding of marking schemes is directly related to the number of raters used in a given assessment episode, these rating schemes are not merely methodological; rather, they are tacitly related to evidence gathering used to draw conclusions about validity and fairness.

Third, the marking schemes shown in table 1.1 reveal research highly restricted in terms of investigating the writing construct. To say writing was undertheorized until the early 1970s (a good milestone date is the 1971 publication of Janet A. Emig's *The Composing Processes of Twelfth Graders*) is an understatement. The sociocognitive models that currently shape the design of writing assessment episodes in the United Kingdom (Weir, Vidakovic, and Galaczi 2013, 212–14) and the United States (Poe, Inoue, and Elliot 2018, 3–38) were not present during the first half-century of our history. As a result of restricted construct representation, table 1.1. may be understood as a taxonomy related to historically embodied forms of evidence. We turn to the use of taxonomies in chapter 10 as the basis of actionable future based on historical patterns.

Fourth, table 1.1 brings forward the importance of genre as constitutive in studying the history of writing assessment (Wood 2018). Popular in the testing of students, these essay-rating methods functioned also as tools of formal research. It worked both ways. Test manufacturers researched scoring methods before making them operational, and those testing methods were borrowed and adapted for independent research, sometimes to investigate the methods themselves. The same method, then, often belongs to two different discourse genres, and the information genre analysis provides, historically, should be interpreted and narrated differently. While ours is not a genre study, the force of genre runs throughout the early history of holistic scoring. In presenting genres of holistic scoring, we see the formation of writing research itself.

MAP 8: FOUR TRADITIONS OF HOLISTIC SCORING

Popular and traditional are not necessarily the same. Tradition implies a transmission and evolution of a practice over generations and, in some cases, across international boundaries. This book discerns and hopes to disambiguate four different traditions of formal holistic essay scoring. They perhaps mark the main departure of our account from previous histories. Map 8 therefore serves as an extension of the continuum of popular essay-scoring procedures shown in table 1.1.

1. **Connoisseurship scoring**. Raters, usually teachers, work from an internalized scale, often a traditional set of academic grades, and assign scores based on their knowledge and experience of students and the educational consequences of the examination. The method may differ little from ordinary teacher grading and thus has deep historical roots. But both in the United Kingdom and the United States, it was used in formal, external examinations for which multiple rating was utilized. Connoisseurship scoring is sometimes described as the inferior form of evaluation that holistic scoring replaced. But it lived on post–WWII in local placement testing, general-college examination boards, teacher-scored exit examinations, and elsewhere. (We take our term *connoisseurship* from the history of the Cambridge English Examinations [Weir, Vidakovic, and Galaczi 2013, 208]).
2. **Trait-informed scoring**. Scorers are trained to focus their reading on a limited set of writing traits (organization, vocabulary, ideas, and so forth) and are sometimes asked to ignore other writing accomplishments. The traits to be used are sometimes defined in a scoring guide or rubric. Like connoisseurship scoring, trait-based scoring has a long tradition, deep rooted in informal classroom practice. It borders on formal analytical scoring methods such as the Diederich scale (1966) or ELL profile scoring (Hamp-Lyons 1987), in which writing traits are scored separately. In the United States the method escalated in popularity during the 1980s and 1990s and is now installed, for instance, in online essay-response schemes, designed for teachers and censured by teachers (e.g., Wilson 2006).
3. **Pooled-rater scoring**. Scorers read papers rapidly—names for the method were "rapid-impression marking" (Britton 1963) or "rapid-impression reading" (Godshalk, Swineford, and Coffman 1966).

Typically, there are three to five independent raters for each essay, and their scores are all accepted, then summed or averaged for a final score. In some forms used in the United Kingdom during the period of our study, post hoc adjustment was undertaken to adjust for consistent variation among raters, and the difference in leniency between the two methods of marking could be standardized afterward. In pooled scoring rater training is usually light, although scorers, usually experienced teachers, may have sample papers to suggest some scale levels or a distribution of scores to shoot for. An experimental testing of the method was reported by British researchers Hartog, Rhodes, and Burt in 1936, and the method was made operational with 11-plus examination essays in Devon by Chief Examiner R. K. Robertson in 1939 (Wiseman 1949). Although it was sporadically used and heavily researched in the United Kingdom into the 1980s, it never caught on. But in the United States, ETS borrowed, tested, and applied it and, eventually, advocated by Fred I. Godshalk and Gertrude Conlan and others, made it the standard scoring procedure for its English Composition Test discussed in chapter 6. Along with the next tradition, pooled-rater scoring stands at the center of the holistic essay-scoring enterprise.

4. **Adjusted-rater scoring**. Typically, each paper has two independent readings, and if the two scores are discrepant, usually more than one point apart, the paper is read a third time, "adjusted," often by the chief readers. Sometimes the method is called "controlled" (e.g., White 1973). Raters are trained with sample papers illustrating the scale and given some sort of scoring guide or rubric to use as they score. The concordance of their scoring with other scorers may be periodically checked by chief readers. As we propose in chapter 5, possibly the first use of this method was in large-scale scoring with undergraduate end-of-course essays graded by the Board of Examiners of the University of Chicago beginning in 1943. The designer was Diederich, who started working for ETS in 1949. Around 1956, adjusted-rater holistic scoring became the standard rating system with Advanced Placement English essays and from there spread around the nation through teachers who had served as AP readers (Advanced Placement Program 1980, 10). In the first half of the 1970s, important high-profile college-essay-assessment programs used adjusted-rater scoring, including the Georgia Regents Testing Program instituted by the Board of Regents of the University of Georgia and the English Equivalency Examination instituted by the California State University system. In the United States, adjusted-rater scoring was probably the most widely used method of holistic scoring.

These four traditions have different provenances, different cadres and followers in practice and research, and different trajectories over the decades. They also ground themselves in different theory and philosophy, as this book hopes to show.

MAP 9: CHAPTERS

We have organized the book into ten chapters. The main narrative chapters, 3 to 9, roughly chronological, focus on particular tests, studies, people, and locales.

- Chapter 2 presents two premises underlying the early historical course of holistic scoring. Social-contagion modeling helps explain the swift and enormous growth of the method in the last half of the twentieth century. Gestalt psychology emphasizes the importance of perception theory in deriving principles behind the dynamic human ordering of the act of holistic scoring. Practitioners of the scoring may have been unaware of these premises, but the principles remain.
- Chapter 3 narrates the attempts at holistic scoring in the United Kingdom up to 1949. It celebrates UK researchers Philip Joseph Hartog and Edmond Cecil Rhodes and educational measurement specialist Cyril L. Burt, who, in their 1936 publication *The Marks of Examiners*, were the first to report and study formal holistic scoring, at least by our definition. The chapter ends with the remarkable use of pooled-rater holistic essay scoring in Devon with government sponsored school examinations, 1939–1948. The history repositions formal holistic scoring within an international context and challenges the received view that the method was invented in the United States.
- Chapter 4 brings the UK history up to the mid-1980s, when it became obvious that in the huge school-leaving and college-matriculation examinations, multiple-marker holistic scoring had lost out to single-marker content-scheme and profile scoring. Despite steady increases in examinees, both first- and second-language students, British researchers and teachers remained faithful to the academic-subject essay and to feedback from examination results to teachers and students. We pay special attention to *Multiple Marking of English Compositions: An Account of an Experiment*, a rigorous study conducted by James N. Britton, Nancy C. Martin, and Harold Rosen into holistic and analytic scoring. The UK assessment experience contrasts with US history, where large-scale distributed assessments largely scored writing competence, not academic-subject knowledge, and where multiple-choice tests of writing gained ground. The chapter concludes with a summary of five major studies, each published in the *annus mirabilis* of 1966.
- Chapter 5 crosses the Atlantic to the United States and explores the role assessment icon Diederich played in devising and installing controlled adjusted-rater holistic scoring, beginning in 1942, through the undergraduate Examining Board at the University of Chicago. In 1949 Diederich moved to ETS, and we pay close attention to changes in early essay assessment of the Advanced Placement Program (1954–1980), changes that solidified the program's widely influential nine-point scale and its method of addressing interrater reliability.

- Chapter 6 features one of the unsung heroes of holistic scoring, Osmond E. Palmer, who directed the Basic College examinations board at Michigan State College for twenty-five years. Palmer was familiar with holistic scoring since he had worked with Diederich at the University of Chicago from 1942 to1946. Years later, in 1960, as chair of the English Composition Test Committee of Examiners, he may have definitively shaped the series of ETS studies that ended with the famous vindication of holistic scoring in Godshalk, Swineford, and Coffman's *The Measurement of Writing Ability* (1966). Palmer is an exemplary early case of the influence of the educational-writing community on the educational-measurement community. Individuals, as we show, can shape industries.
- Chapter 7 turns away from postsecondary education to a history of holistic scoring in the California schools from 1960 to 1982. The main story recounts a resistance of school teachers to mandated state accountability testing. We feature unexamined work of Albert "Cap" Lavin at Sir Frances Drake High School and of Catherine Keech with the Bay Area Writing Project. In exploring tensions arising around the first assessment of teacher-led curricular initiatives by Michael Scriven, we come to see the value of resistance to restrictive forms of accountability. As we claim, the integrative processes of teaching and assessing writing established in California remain in schools and colleges across the United States.
- Chapter 8 reviews the contributions of independent academic-assessment researchers to the early support and critique of holistic scoring. It continues with a narration of another independent study of writing assessment, the first assessments conducted by the National Assessment of Educational Progress (NAEP) between 1974 and 1984, featuring the friction between holistic essay scoring and primary-trait scoring and between early Education Commission of the States approaches to assessment and those of ETS, who administered the 1984 assessment. The chapter explores the fluid research relationships that arose between independent researchers and large-scale testing organizations in the creation of a body of knowledge for writing studies that relied on holistic scoring. The chapter closes with a recollection of the work of Rexford Brown and his role in the early years of NAEP.
- Chapter 9 ends our histories of holistic scoring by focusing on the inferences, interpretations, and uses of three artifacts of holistic scoring: splitters (writing samples associated with failure in interrater reliability), rubrics (scoring methods utilizing a checklist of selected writing traits), and profiles (outcomes reported to students in the form of separate trait scores). In doing so the chapter identifies the limits of holistic scoring associated with score interpretation and use and pays special attention to Liz Hamp-Lyons's 1987 dissertation, which argued, in effect, for abandonment of holistic scoring.

- In chapter 10, we explore the possibility of actionable history. Continuing our attention to validity, reliability, and fairness, we reanalyze the major studies conducted from the mid-1930s to the mid-1980s using a category-of-evidence (CoE) framework. As we argue, rapid demographic shifts over the next forty years necessitate new genres of assessment in which lessons learned from the early history of holistic scoring return to play an important role in our common future.

These chapter maps are intended as interpretative guides to our historical recovery project. Our intention in providing them is to prevent closure and encourage further historical study of writing assessment. The history we now present is, as we say, more mosaic than fresco, dwelling in depth on a few places, stories, and documents (such as those identified in tables 10.1 and 10.2). Tesserae can be found in the over 1,000 annotations in our accompanying WPA-CompPile Research Bibliography, No. 27 (Haswell and Elliot 2019). While we trust our history provides insight on the origin, development, and significance of the assessment genre of holistic scoring, much work remains that can be informed by the maps provided here.

NOTES

1. T. J. Elliott, Chief Learning Officer at the Educational Testing Service, wrote to us that above all, readers of this book should understand the "hybrid and even variegated nature of holistic scoring," that "there is no 'sure' or standard version" (pers. comm., May 2018).
2. Today US students spend up to six weeks of the academic year preparing and sitting for mandated tests (Nelson 2013, 3). A study sponsored by the Center for American Progress found that "students take as many as 20 standardized assessments per year and an average of 10 tests in grades 3–8," with "urban high school students spend[ing] 266 percent more time taking district-level examinations than their suburban counterparts" (Lazarín 2014, 3–4). Those of us who were at our school desks in the 1950s and 1960s remember nothing like that fixation.
3. In order of start-up: *Studies in Educational Evaluation* (1974), *Assessment and Evaluation in Higher Education* (1975), *Evaluation Review* (1977), *Notes from the National Testing Network in Writing* (1982, ceased in 1990), *Language Testing* (1984), *PARE: Practical Assessment, Research & Evaluation* (1988), *Assessing Writing* (1994), *Educational Assessment* (1995). Today we can add three more to the list: *The Journal of Writing Assessment* (2003), *Research and Practice in Assessment* (2006), and *The Journal of Writing Analytics* (2017).
4. In essence, history is not what happened in the past because, as Sartre's protagonist Roquentin famously puts it, "Le passé n'existe pas" (the past no longer exists). As we argue in chapter 10, the past is interpretatively fluid.
5. White borrowed his four overarching emplotments from the literary theory of Northrop Frye—a reasonable source given White's insistence that history is an imaginative construct.
6. It may seem that the term *trajectory*, with its common meaning as the path of an energized physical object, is inappropriate for our application to historicizing. But

the trajectory of a fired missile is only part of its story. Why was it fired and aimed in the first place, and what will be the outcome? Note that Hayden White (1980) defined "chronicle" as a genre of historytelling that "lacks closure," does not provide the "summing up of the 'meaning' of the chain of events" that one expects in a "well-made story" (20). White's "chronicle" and our *trajectory* are roughly synonymous.

7. Claims such as these are common: ETS "laid the foundations for holistic scoring" (Burstein, Leacock, and Schwartz 2001); ETS was "the originator of holistic scoring" (White 1993, 82).
8. Over time, one of the most counterproductive instances of polysemy has been interrater reliability. The term can refer to statistical methods that, in fact, calculate reliability estimates quite differently and point toward opposing premises in theory (for critique, see Cherry and Meyer 1993).
9. The genre of the essay is used throughout this history because it is that form that is most identified from the mid-1930s to the mid-1980s. We recognize that the essay is not merely a form of writing but, rather, is part of traditions of artifact production, use, and interpretation—ideologies that shape the very contexts in which they emerge (Elliot 2016; Gee 2012; Miller 1984; Spinuzzi 2003, 2015; Wood 2018). Readers may well wonder how diverse genres and stakeholders have shaped the history we present—and how the future of teaching and assessing writing will be shaped by broader representation.