Assessing Writing 29 (2016) A1–A5

Contents lists available at ScienceDirect

Assessing Writing

Editorial

Farewell to holistic scoring. Part Two: Why build a house with only one brick?

In Part 1 of this Editorial (*Assessing Writing* January 2016) I began my call for the "end of holistic scoring", and opened an argument for what I call multiple trait scoring. In fact, I am not arguing against holistic scoring so much as arguing for a fresh and closer look at it together with the other options now available to us for carrying out the important work of making fair decisions about the quality of written work, especially in high stakes contexts.

I discussed holistic scoring in Part 1 of this Editorial; let me now turn to the challengers to holistic scoring: analytic marking, impression scoring, and multiple trait scoring. Analytic scoring is often confused with or assumed to be synonymous with multiple trait scoring, but multiple trait scoring is *not* the same thing as 'analytic scoring'.

'Analytic scoring' has its origins in the early days of psychometrics, and gained strength at a time when educational measurement practitioners in the US were scrambling to find ways to identify large numbers of suitable candidates for the military as that country prepared to join the First World War. They needed something quick and wanted it to be reliable; and their focus was on testing intelligence. The chosen solution was the Army Alpha tests (Yerkes, Bridges, & Hardwick, 1915). Spolsky (1995), using primary and later sources, reports on this development and comments that "with all their weaknesses and confusion (they) were the beginning of mass objective testing of mental abilities" (p. 36); and identifies this as the first introduction of what we now call 'multiple choice' test items. The Alpha tests were based on analytic/frequency measures such as average word complexity, syntactic complexity, use of idiom, etc. (here Spolsky cites Henmon, 1929), and are best related not to writing assessment but to the focus of these early psychometricians on the measurement of intelligence, with language as the quantifiable aspect of intelligence.

Just after World War 1 analytic and objective 'measures' of intelligence and of language skills were enthusiastically adopted, mainly for foreign languages testing in the beginning, but by1926 were being used for the College Board's Scholastic Aptitude Test (Spolsky, 1995). Not surprising then, that as American colleges began to grapple with the difficulties of judging large quantities of written work, the bureaucrats pressed for these 'new-type' "objective" tests which could be rapidly scored analytically. But not surprising either that classroom teachers of writing resisted this movement and argued – persistently and for many years, if initially not very successfully, that real writing can only be judged by real readers reading whole texts (Fader, 1986; Myers, 1980).

'Analytic' scoring was an attempt to attend to and value specific, explicit aspects of forms of human behaviour that can be (or at least, were and are claimed to be) occurring simultaneously. 'Analytic' marking consisted (and in some places still consists) of a list of elements felt to be important in written prose, each of which was labelled with some pseudo-linguistic term, and a single number reported for each element. Usually the elements were listed separately—a categorization necessary for the factor analysis methods then coming into common use—but then totaled and reported as a single number, often out of 100. Some elements might be 'weighted', i.e., judged to be more important, and doubled or tripled in 'value'. Appendix 1 shows the scale created and used by Diederich, French, and Carlton (1961). As we can see, Diederich et al. valued such curious factors as *movement* (within a single element comprising organization, relevance, and movement), *flavour* and *individuality* (these two also considered as a single element), and *punctuation* (a single element, but happily for many students, weighted at five times less than organization and its bedfellows!).

At that time the opposing view, though existing in many variations, was what was generally known as 'impression marking' or impression scoring. Impression marking was the choice made for exams in England during that period; it continues to be practiced in many school-level tests of English 'language'/writing, and is still often used in mainstream classrooms for marking writing in low-stakes, i.e., single classroom, internal, contexts; and it is used in periphery education systems where a single marker, perhaps with occasional sample second-marking are all the resources available—a practice all readers of this journal will agree is regrettable, at best. As Weir, Vidaković, and Galaczi (2013) remind us, even single

http://dx.doi.org/10.1016/j.asw.2016.06.006

impression marking with very careful briefing of markers and moderation, has been shown to be less reliable than having two or more markers (pp. 200–201). In many institutions around the world, the assessment of the English writing of second language writers is done by impression marking with two markers ("double marking"). Wiseman (1949) claimed that double marking produced significantly improved reliability, but later study replications have not always found similarly good reliability (Ofqual, 2014); and of course, the fact that these scores are impressionistic means there is no way to uncover why the reliability has been found to be poor.

As I interpret the history, the 'impression' approach to judging a text developed within the long tradition of writing instruction from Quintilian through the medieval European universities such as Bologna and Paris, to the increasingly frequent written exercises and themes that accompanied and gradually supplanted the oral disputations at Oxford and Cambridge (and at Harvard and other Ivy League US universities) during the first half of the nineteenth century (Montgomery, 1965), and which were dominant in the British 'grammar schools' from the latter third of the nineteenth century (Ferreira-Buckley & Horner, 2001). I can personally testify that this pedagogic approach still continued when I was at grammar school in the late 1950s and early 1960s.

Berlin (1987) describes something similar as being the teaching approach in the first half of the twentieth century at Harvard, perhaps the most prestigious university in the US; he perhaps ironically suggests it was "designed to provide the new middle-class with the tools to avoid embarrassing themselves in print" (p. 35):

> "students wrote a theme for each class day on uniform theme paper. The choice of subject for four of these was limited to descriptions of surrounding scenes, while the other two required a translation from Latin, Greek, German or French (normally on Saturdays), and a summary or comment on the lecture of the period when all sections met together. All themes were read and corrected with the use of an abbreviated set of marks of correction. The emphasis in this method was on superficial correctness—spelling, punctuation, usage, syntax—and on paragraph structure. Students were often asked to rewrite themes with a view to correcting their errors." (Berlin: 37–38)

We have here a combination of error correction and impression marking; this awkward partnership did not fit well with reforming compositionists in the US who hoped for a more humanist/expressivist composition curriculum (for example, Elbow, 1973; Murray, 1968).

It is worth nothing that this approach had never been adopted as the approach of the British secondary schooling system (that is, the vast majority of schools outside the so-called 'public' schools to which only the elite, and predominantly males, went). In Britain, 'marking' meant actually making marks onto student work, and the marks were not numbers alone: typically, marks related directly to comments written by the teacher (Britton, Rosen, & Martin, 1966; LATE, 1965; Wilkinson, Barnsley, Hanna, & Swan, 1980). Experiments with analytic scoring quickly generated concern and rebellion, as the liberal '60s brought reform in many areas of education, and gave confidence to English teachers, led by influential teacher educators such as James Britton, a driving force for the founding of the London Association of Teachers of English in 1947, and James Moffet (1968) to resist attempts to bring so-called objective tests such as multiple-choice tests of 'grammar', into English examinations. Teachers' groups and teacher-educators began to express concern about what is lost or left behind when 'quality' is constrained as a single number. As Shaw and Weir show, the US tradition of psychometrics never caught on in the UK.

In the US, the judging of student writing was, then, part of a tradition that was quite conflicted, emphasizing as it did 'textual appreciation' through great works, first of the Classics and later of English literature, while attempting to teach students to write well by emphasizing form (sentence structure, grammar, word choice, spelling and punctuation) using the tradition of teaching Latin grammar, and further compounded by the psychometric community which was then taking charge of what counted as evidence of student learning. How much more conflicted and confusing, then, when looked at internationally. From the 1960s on, teachers, scholars and researchers in many countries have been influenced by one or more of the traditions in judging writing as they have confronted the issues of responding to student work and reporting on it as a measured quality. As demands to assess, and report judgements of, actual writing grow, as student numbers grow near-exponentially, and as test candidates, their parents, their future teachers, and maybe even their employers expect to understand and critique the judgements made about them, it has become clearer and clearer that what is needed is not mere numbers, and certainly not just *one* number, but something far richer. In fact, for many teachers of the 1970s era it had become clear before the end of that decade that to ensure meaningful assessment of writing, and to convince higher authority of the value-added of direct writing assessments to education and the well-being of society as a whole, it was important to *both* argue for validity of writing assessments *and* demonstrate reliability of these assessments across time and circumstance.

The felt need for a single assessment instrument that could show both high reliability and explicit validity, and that could make sense to all stakeholders in an assessment context were very important reasons for the emergence of trait-based approaches to judging writing. But another important reason for the move in this direction has been to bring assessments back into the hands of teachers.

Moving from analytic to multiple trait scoring

In 1981, Jacobs, Zingraf, Warmuth, Hartfiel, and Hughey, teaching at Texas A&M University, published their seminal English Composition Program, a series of three books, on teaching, learning and testing composition respectively: their ESL