Dr. Les Perelman critiques the current NAPLAN writing task and argues Australia can make a great essay assessment…

Introduction

State and national educational testing is common throughout most of the world, although its uses vary. Because writing is such a primary and essential ability, it is almost always included in any largescale educational assessment. This report has four major purposes. First, to review briefly the essential concepts underlying validity in writing assessments. Second, to review interesting and differing approaches to essay assessment in Anglophone countries. Third, to discuss the writing assessment on the National Assessment Program Literacy and Numeracy (NAPLAN) in terms of its stated goals, its design, and contemporary validity theory. Finally, the report will present some possible suggestions for developing a new NAPLAN writing assessment that would better fulfil one or two of its articulated functions and better promote classroom learning.

The Executive Summary is available in the full report.

Contemporary concepts of validity

Traditionally, the validity of a test was based on three interrelated concepts that are often best framed as questions. First, construct validity is concerned that the assessment instrument is measuring the abstract ability of interest, the construct, and that the theoretical basis of the construct is sound. In order to gather evidence related to the construct under examination, the construct needs to be defined, and observable variables that represent the construct need to be specified. In the case of NAPLAN, for example, the specific Australian Curriculum objectives involving writing ability help to define the construct. Construct validity also asks whether the instrument measures features that are irrelevant to the construct. Eliminating construct-irrelevant variance is thus essential to a well-defined assessment.

A second facet of the traditional validity framework, content validity, also is concerned whether the measure adequately covers the domain of abilities that constitute the construct. A third facet of validity calls for various types of external or criterion validity. Does the assessment instrument predict performance on other measures that substantially incorporate the construct? Does it adequately predict future performance in activities closely related to the construct? This threefold view of validity — often referenced as the Trinitarian model — was first introduced in the 1966 edition of the Standards for Educational and Psychological Tests and Manuals (American Psychological Association, 1966). In the following half century, American psychometricians have reframed and expanded the notion of validity to focus on the interpretation and uses of scores. This view culminated in the 2014 edition of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, et al, 2014). A measure is not valid in itself but only in terms of how it is interpreted and how it will be used. Here is how Michael T. Kane, a member of the joint committee that developed Standards, framed the relationship between validity evidence and the consequences of score use:

> "In addition to their use in making decisions about individual test takers, tests also have been used for policy analysis, program evaluation, research, and educational accountability; in all of these cases, the requirements imposed by the intended use shape the design and development (or the selection) of the tests. Concerns about validity have their roots in public concerns about the appropriateness, or fitness, of the test scores for their intended use or uses." (Kane, 2013)

Consideration must be given to the potential risks of a measure's misuse. There often are, in particular, issues of fairness that need to be carefully considered. For example, are there construct-irrelevant features in the test that might penalise some groups while favouring others? A classic example is a test item from an Scholastic Aptitude Test (SAT) analogies section from the early 1980s.

RUNNER: MARATHON

a) envoy: embassy
b) martyr: massacre
c) oarsman: regatta *the correct answer*
d) referee: tournament
e) horse: stable.

On this question, 53 per cent of whites but just 22 per cent of African Americans chose answer C (John Weiss, 1987). The reason is obvious; most inner-city African-American students, as well as, probably, most students growing up in the Australian outback, probably have neither participated in nor watched many regattas. Yet knowledge of upper class aquatic sports is irrelevant to assessing the ability to perceive abstract analogies.

It is also clear that tests have unintended and sometimes harmful consequences. Various studies of provincial school literacy assessments in Canada, for example, have indicated that these assessments narrowed what was taught in classes by an emphasis on reductive exercises, as well as overreliance on test preparation at the expense of teaching invention and revision (Cheng, Fox and Sheng, 2007; Luce-Kapler and Klinger, 2005; Slomp, 2008). Slomp, in particular, recounts the demoralising effect that narrowly focused large-scale assessments can have on both teacher morale and curricular effectiveness.

Issues of ethical use, fairness, and attention to possible consequences, along with other considerations, have caused American psychometricians to abandon an objectivist view of validity and, instead, to focus on socio-cognitive orientations of validity (Mislevy, 2018). Rather than viewing validity as a property of the test, psychometricians now conceive validity to be a property of the proposed interpretations and uses of the test score. Consequently, validation is a two-fold argumentative process concerning, first, the interpretation of the test data and, second, its use. Messick describes the argumentative nature in detail in his chapter on "Validity" in the third edition of Educational Measurement (Messick, 1989b). In the fourth edition of that work, Kane's chapter on "Validation" expands on Messick and presents a very detailed explanation of the twin concepts of validity and validation (Kane, 2006). Kane defines validation as "the process of evaluating the plausibility of proposed interpretations and uses" and validity "as the extent to which the evidence supports or refut**es the proposed interpretations** and uses" (p17). From these definitions come two kinds of arguments: the interpretative argument and the use argument. "An interpretative argument specifies the proposed interpretations and uses of test results by laying out the network of inferences." Recently, Kane (2013) emphasises the importance of the consideration of use in determining validity. Consequently, Kane now reframes the process as a single interpretation/use argument.

Types of assessments

One essential difference between writing assessment and most other academic assessments is that writing is a skill rather than a body of knowledge (Elliot, 2005). Although writing embodies components that consist of bodies of knowledge, such as spelling, as a whole, it resembles engineering more than it does physics or history (Perelman, 1999). The essence of writing is constructing an artefact to be used, that is to be read, by someone. While language itself, as manifested in speech, is an innate human feature (Berwick, Friederici, Chomsky and Bolhuis, 2013), writing is a technology that is younger than fire, having existed for only four to five thousand years (Fischer, 2001). With writing there is no correct answer. In mathematics there may be several ways to prove theorems or solve equations, but the number is always finite. The permutations of possible written texts are infinite. It is this characteristic that makes constructed response assessments, that is writing tests, the only avenue for authentic evaluation.

Purposes of writing assessments

A feature that defines any writing assessment is its purpose. The discussion on validity tells us that a test's use defines any process of validation. To ignore a test's use in its design or to employ a test designed for one function to serve an entirely different purpose is highly problematic. As Ananda and Rabinowitz note:

> Some high-stakes assessment systems try to create efficiencies by using the same test to serve multiple purposes and draw multiple inferences. This is a questionable practice. For example, many tests that are incorporated into new statewide student assessment systems were designed to assess a student's content knowledge or to assess student achievement relative to a norm group of students at a particular point in time (e.g. Stanford Achievement Test-9, Terra Nova). However, some states are also using these tests to help determine whether schools and teachers should be rewarded or sanctioned — a purpose for which the tests were not designed. (Ananda and Rabinowitz, 2000)

Writing tests are used for a variety of purposes. They are used to place students in specific classes. They are used to provide feedback to students and parents to indicate areas of strength and weakness as well as to situate a student's performance on the test relative to some larger cohort. They can be employed by a teacher as valuable information to identify specific students' strengths and weaknesses. They are used to assess the overall performance of schools, districts, states and nations. The performance of students on such standardised tests can also be used by parents to select schools for their children. In some educational systems, students are required to pass a writing examination to graduate from secondary schools. In many countries, some sort of writing examination is included in the list of tests students take for admission to tertiary institutions.

Different uses demand different types of writing assessments with different writing conditions, different writing media, different types of prompts, different genres of writing, different personnel marking the assessment, and different methods of marking. These variables all need to be considered in terms of the objective or objectives of the assessment.

Writing conditions

Where and how students write scripts for assessment are two key considerations. If the objective of the assessment is to provide formative feedback to the student, the assessment may very well be done at home or online. Assessments can also be done in the classroom, but time becomes a major factor of what is assessed for any purpose. While short writing assignments of 20-45 minutes are able to produce reliable agreement among markers (Godshalk, Swineford and Coffman, 1966; White, 1984, 1994), much of the score (and agreement) reflects essay length: the longer the essay, the higher the mark (Perelman, 2012). These short, timed tests may be able to assess some elements of the writing construct such as fluency, but other important elements of the construct, such as planning and revising, are absent (Deane, 2013). Longer periods give students some opportunity for planning and revision, but not as much as an assignment completed over one or more days. These last scenarios, however, introduce issues of test security.

Writing media

Writing, in its history, has employed clay tablets, papyrus, parchment, and paper as well as metal styli, quills, fountain and ball-point pens, typewriters, and computers. Currently, the two major forms of inscription are pen or pencil on paper and computer writing. There is also texting and its related forms, but because of the extreme limitations of length, they will not be considered. Writing on a computer is substantially different than writing with pen and paper, and although the literature reports substantial benefits for classroom use (Cochran-Smith, 1991; Lam and Pennington, 1995; Liu, Moore, Graham and Lee, 2002; Robinson-Staveley and Cooper, 1990), the use of computers in extended constructed response tests appears to privilege students with fast keyboarding speed and to penalise students with slow keyboarding skills (Russell, 1999).

Writing prompts

Writing prompts can differ both in the specificity of the instructions for the essay and in the material provided to the student. Prompts can be very specific about content, form, or both. The prompt can include readings, charts, and graphs, or it can include a simple generic prompt, such as "Is failure necessary for success?", intended to be answerable by anyone in the test population. It can also be based on set readings given out before the examination. The instructions in the prompt can also require that the writer take a specific stance or specify a specific structure. It can also indicate to the student what features are the most important or should be emphasised.

Writing genres

Although there are various genres and purposes for writing, the three most common genres for assessment in primary and secondary schools are narrative, informative, and persuasive or in the taskbased vocabulary of the United States' National Assessment of Educational Progress (NAEP): to convey experience (real or imagined); to explain; and to persuade (National Assessment Governing Board, 2010; National Center for Education Statistics, 2016; Persky, 2012). Often the choice or mix of prompts is determined by year, with the lower grades favouring the narrative genre and the upper grades favouring persuasive writing. The National Assessment of Educational Progress (NAEP) assessments, for example, administer two 30-minute prompts to selected Grade 4, 8, and 12 populations. As displayed in Table 1, the mix, however, changes, with conveying experience (narrative) constituting 35 per cent of the prompts in Grade 4 and decreasing to only 20 per cent in Grade 12, while the upper grades emphasise informative and persuasive writing.

This age-based progression of emphasis from narrative to informative and persuasive is common to many Anglophone writing assessments.

Table 1: NAEP Percentage Distribution of Communicative Purposes by Grade

Purpose	Grade 4	Grade 8	Grade 12
To Persuade	30%	35%	40%
To Explain	35%	35%	40%
To Convey Experience	35%	30%	20%

Source: National Assessment Governing Board, 2010

Evaluators

Another important variable in any assessment is the proficiency and training of the markers. Some assessments have local teachers mark the essays in nearby venues. In other cases, there are large regional marking centres where teachers are first trained and then participate in marking sessions. During the past 10 to 15 years, some major tests have had evaluators mark assessments online at home. In some cases, individuals train at a regional centre and then mark at home. In other cases, both the training and marking occur online.
Although many writing assessments use only teachers certified in English, this practice is not always the case. For example, Pearson Education, a large multinational corporation that held in 2013 the contract to construct, administer, and mark the Texas state writing tests, advertised on Craigslist for markers holding a university degree in any field, offering the rate of USD12 per hour (Strauss, 2013).

Another crucial factor affecting evaluators is the amount of time budgeted per script. Even in scoring sessions in which there is no explicit quota for the number of scripts per hour, there is always an approximate amount or range of hours budgeted for the entire marking session or sessions. When marking is outsourced to commercial companies, especially for very large-scale tests, the marking quotas sometimes become draconian. The private companies conducting the grading of the now defunct SAT Writing Test Essay required markers to mark between 20-30 scripts per hour or one script every two to three minutes (FairTest, 2007; Malady, 2013; Joanna Weiss, 2014).

Marking methods

There are three principal methods of marking writing tests, although one is rarely used and another has several variations: multi-trait or analytic scoring, holistic scoring and primary-trait scoring.

Multi-trait or analytic scoring

Various forms of multi-trait scoring have been long used in classrooms in formative assessments meant to identify areas of strength and weakness. Analytic scales have also been common in UK examinations, usually focusing on two to four traits marked on multiples of a five-point (level) rubric. In the US in the late 1950s, researchers at the Educational Testing Service recruited 53 readers from various professions, including journalism, law, and business, to judge 300 papers written by college freshmen. The readers ranked the overall quality of the essays in nine subcategories, applying three gradations of high, medium, and low. A factor analysis of the readers' comments identified five factors (Diederich, French and Carlton, 1961):

1. Ideas (relevance, clarity, quantity, development, persuasiveness)
2. Form (organisation and analysis)
3. Flavour (style, interest, sincerity)
4. Mechanics (specific errors in grammar, punctuation, etc)
5. Wording (choice and arrangement of words).

Subsequent formulations usually added a category concerned with sentence structure, fluency and style, producing a six-trait model, and some, usually concerned with handwritintg in the lower grades, added a seventh trait focusing on presentation (Bellamy, 2001; Murray, 1982; Purves, 1992; M.A. Smith and Swain, 2017; Swain and LeMahieu, 2012).

The five traits described above appeared to adequately represent a consensus construct of writing ability, but the researchers discovered that it was difficult to achieve adequate inter-marker reliability. Recently, there have been successful undertakings to make multi-trait scoring more reliable while retaining the valuable information it can provide. The most successful has been the Analytic Writing Continuum of the US National Writing Project (Singer and LeMahieu, 2011; M.A. Smith and Swain, 2017; Swain and LeMahieu, 2012). The editor of the journal Assessing Writing has written two eloquent editorials arguing for the pedagogical and programmatic priorities inherent in developing and employing sound analytically scored tests. [Full disclosure: I am a member of the editorial board of this journal.] Some multi-trait models, especially ones that are reporting just a few traits, report each trait individually. The National Writing Project's Analytic Writing Continuum adds a holistic score, which is determined first by markers, in addition to six slightly reformulated analytic scores (Singer and LeMahieu, 2011; Swain and LeMahieu, 2012).

A feature common to almost all multi-trait scoring schemes is that they maintain identical or equivalent scales for each trait. Some traits may have larger scales, but they are always integer multiples of the base scale, allowing markers to always use the base scale for the primary determination and then, in the case of the larger scores, adjust the marks within the range of the multiple. If the base scale is 1-5, for example, and some traits are on a 1-15 scale (multiple of 3), markers first determine a level of 1-5 and then refine that score by determining if it is high, middle, or low. A level 4 score, for example, could be 12 (middle), 11 (low), or 13 (high).

Many tests determine a final composite score by summing the trait marks. In some cases, some marks are given additional weight in determination of the composite score by doubling their value. In other cases, the individual traits are simply summed. In all of the cases in which a composite score is employed, there needs to be explicit validity arguments, including those on interpretation and use, which demonstrate:

1. that the specific construct of writing ability is being adequately represented by the sum of sub-scores; and
2. that the test results are being used appropriately and fairly. The issue of construct representation is of particular importance in a composite score. Evidence needs to be presented that the traits model the construct of interest in the correct proportions.

Holistic scoring

Holistic scoring was developed as a method to achieve much greater reliability than multi-trait scoring, which often had large inconsistency among markers' scores. Around the same time as the factor analysis that resulted in the analytic categories discussed above, another team of researchers at the Educational Testing Service (ETS) were trying to solve the problem of inter-marker reliability. They did discover that they could achieve acceptable levels of inter-rater reliability, correlations of 0.7 to 0.8, by training readers to rate essays holistically (Elliot, 2005; Godshalk, Swineford and Coffman, 1966; White, 1984). The basic premise underlying holistic reading is that the whole is greater than the sum of its parts, especially when a mind confronts a written text. Readers do not count mistakes, although mistakes can certainly impede reading; they seek meaning, along with some sort of efficiency and, if possible, some elegance and beauty. Readers do not count the number of paragraphs or the number of sentences in a paragraph; they care if the text is complete, informative, and compelling. If it is a onesided conversation in which all of the reader's unstated questions, comments, and objections have been addressed, the text is successful (Pratt, 1977). If it does not do these things, it is not. A holistic scale measures the relative success of a text but does so through a rubric that incorporates many of the traits in analytic scoring as heuristics towards a conception of a whole rather than as a sum of autonomous components.

Primary Trait Scoring

Primary trait scoring is similar to holistic scoring in that markers produce a single mark. However, rather than representing an evaluation of the entire script, a primary trait score evaluates the single trait of concern, for example, persuasiveness. After its initial development, it was employed briefly by the National Assessment of Educational Progress (Lloyd-Jones, 1977; Persky, 2012). This marking technique can yield very precise and reliable information on a single component of the writing construct or student performance in writing a specific genre. However, because the information it yields is so narrow while the technique requires significant resources for training markers, it is rarely employed.

Comparative Anglophone school writing assessments

Any evaluation of the NAPLAN Writing Assessment would be incomplete without comparing it to other writing assessments. Six other assessments were chosen to represent different approaches and purposes in Australia, the United States, Canada, and the United Kingdom. Three of these assessments are for primary and secondary years and have similar purposes to those of NAPLAN. The other essay tests are for Year 12 students and are a component of university entrance qualifications. All of the assessments were chosen because, in addition, they displayed different prompt types, different essay genres, and different marking schemes. They are meant as comparisons, not in any way as a representative sample. Table 2 summarises some of the features of these seven tests. The discussion below will summarise the relevant and salient features of the six tests exclusive of NAPLAN. The following section will then discuss the NAPLAN essay in detail using, at times, these six other writing tests as points of comparison.

The Comparative Anglophone School Writing Assessments section is available in the full report

The Australian literacy curriculum, NAPLAN, and the writing construct

The National Assessment Program Literacy and Numeracy (NAPLAN) was established in 2008 by then education minister Julia Gillard. NAPLAN has three explicit purposes:

1. as an indicator of school performance for use by school administrators in improving schools and by parents in selecting schools
2. as a snapshot of national student achievement that can be used for longitudinal comparison
3. as formative feedback to students on their skills in literacy and numeracy (Australian Curriculum Assessment and Reporting Authority, 2017b). In terms of the NAPLAN writing essay, underlying all those purposes is the crucial assumption that the essay elicits, and the scoring measures, a reasonable approximation of a general writing construct.

A major problem in evaluating the NAPLAN writing essay is that, unlike most other Anglophone largescale writing assessments, there are no publicly available specification or framework documents. When considering the design and implementation of the test, we do not have access to justifications or explanations for specific choices. Consequently, the discussion of efficacy of the writing essay will begin by reviewing the relevant curricular goals and then discussing the test's design, with particular attention to some of its more unusual elements and design choices. A detailed discussion of the marking criteria for some of the NAPLAN essay's 10 traits will then be followed by an examination of the top-scoring training scripts for the persuasive essay.

The Australian Curriculum writing objectives

The Australian Curriculum lists two primary objectives directly connected to writing:

- Plan, draft and publish imaginative, informative and persuasive texts demonstrating increasing control over text structures and language features and selecting print, and multimodal elements appropriate to the audience and purpose (ACELY1682)
- Re-read and edit texts for meaning, appropriate structure, grammatical choices and punctuation (ACELY1683).

In terms of the relevant portions of the published ACARA Achievement Objectives for the Years of NAPLAN tests, the specific goals are:

- Year 3 — Students create a range of texts for familiar and unfamiliar audiences. They demonstrate understanding of grammar and choose vocabulary and punctuation appropriate to the purpose and context of their writing. They use knowledge of letter-sound relationships including consonant and vowel clusters and high-frequency words to spell words accurately. They re-read and edit their writing, checking their work for appropriate vocabulary, structure and meaning. They write using joined letters that are accurately formed and consistent in size.
- Year 5 — Students create imaginative, informative and persuasive texts for different purposes and audiences. When writing, they demonstrate understanding of grammar using a variety of sentence types. They select specific vocabulary and use accurate spelling and punctuation. They edit their work for cohesive structure and meaning.
- Year 7 — Students create structured and coherent texts for a range of purposes and audiences. When creating and editing texts they demonstrate understanding of grammar, use a variety of more specialised vocabulary and accurate spelling and punctuation.
- Year 9 — Students create texts that respond to issues, interpreting and integrating ideas from other texts. They edit for effect, selecting vocabulary and grammar that contribute to the precision and persuasiveness of texts and using accurate spelling and punctuation. (Australian Curriculum Assessment and Reporting Authority, nd)

These objectives, although important, are extremely limited and highly reductive, placing mechanical correctness in the foreground while giving some, but limited, acknowledgement of writing as primarily a communicative act. First, there appears to be a strong emphasis on spelling. There is, however, some irony in this weighting. Incorrect spelling as a common problem in written discourse only occurs in languages such as English, where for various reasons — in the case of English mainly historical — there are often disjunctions between phonology and orthography, that is between pronunciation and spelling. In languages such as Spanish, for example, where the relation between the two is much closer, spelling is much less important for the teaching of writing. The irony in Australian education exists in the recent attempt by the Australian Federal Government to institute phonics testing in Year 1, including having students sound out nonsense words to test "phonic awareness" (Ireland, 2017). Yet, if a Year 3 student employs that same phonic awareness to write "nite" instead of "night", he or she is penalised.

There is further irony in this emphasis on correct spelling because of ACARA's desire to move NAPLAN online. As mentioned previously, similar online tests allow students to employ the word-processing applications, including a spell-checker. In designing an online system with word-processing features, ACARA probably had to modify it by removing an already existing spell-checker. The ubiquity of these word processing applications has significantly reduced spelling errors in student writing. In 1986, a study of errors in a stratified representative sample of 3000 first-year American university essays revealed that spelling errors accounted for more than 30 per cent of all errors identified in the papers (Connors and Lunsford, 1988). When the study was replicated 20 years later in 2006, spelling errors (including homonyms) constituted only 6.5 per cent of all errors (Lunsford and Lunsford, 2008). Just as the electric typewriter reduced the importance of handwriting and made texts more legible, the spell-checker assists students in spelling correctly.

Implied in ACARA's progression of writing competencies, appears to be some premium on the use of specialised and, possibly, uncommon vocabulary. This tendency will become even clearer in the examination of the marking rubric. However, the use of multi-syllable, less frequently used language, unless it is necessary for precision of meaning, in place of plainer more accessible language goes against the received wisdom of such great 20th century Anglo-American stylists as Pinker (2014) Struck and White (1979), Gowers (1973), and Orwell (1954).

Although there are frequent references to editing, most frequently for grammatical correctness, there are no references to revision. For the past half-century, composition theorists and researchers, especially those in the US have emphasised the difference between editing, which is concerned with grammatical correctness and specific vocabulary choices, and revision, which is literally the re-seeing and reformulating of ideas by adding, subtracting, elaborating, and explaining, among other actions. Because research has demonstrated that the ability to revise effectively differentiates mature writers from novice writers (Flower and Hayes, 1981; Sommers, 1980, 1981, 1982), teaching revision is an essential component of teaching writing (Adler-Kassner and Wardle, 2016; Grabe and Kaplan, 1996; Murray, 1982; Newkirk, 1981) and ignoring it, ignores a major part of the writing construct.

Finally, instead of embracing the essential connection between reading and writing, the ACARA Achievement Standards segregate achievement into two modes: receptive and productive. As powerfully articulated by several of the test specification and framework documents referenced earlier, reading and writing, reception and production are intertwined. A child acquires language without explicit instruction through the exposure of an innate biological language facility to a specific human language. To be more direct, a child learns to speak a language by being spoken to. Reading strongly influences the form and substance of a child's writing, and writing allows a child to appreciate the choices made in a written text and understand it more completely.

Design

The NAPLAN essay follows a section on Language Conventions which contains approximately 40-60 items concerned mainly with spelling and pronoun and article usage. Until 2014, the same prompt was given to Year 3, 5, 7, and 9 students. Beginning in 2015, one prompt was given to Year 3 and 5 students and a different prompt to Year 7 and 9 students. The essays are scored without the marker knowing a student's year. Beginning in 2018, some students will be writing online, producing printed text. Because no Year 3 students will be writing online, markers will be able to identify Year 5 students writing online. Each year, all students write in the same genre, either persuasive or narrative. Since 2010, there have been two years with narrative prompts (2010 and 2016) and six years with persuasive prompts (2011, 2012, 2013, 2014, 2015, and 2017). The prompts from 20112016 are displayed in Appendix I.

Curiously, even though the Australian Writing Objectives specifically include informational writing, that genre for some unexplained reason has been excluded from NAPLAN. As evidenced in the review of other Anglophone state and national writing tests, informational writing is a common a genre in school writing assessments. In school and in the workforce, individuals will probably write more informational texts than persuasive texts and certainly more than imaginative narratives. Indeed, although this report contains some persuasive elements, its primary genre, at least so far, has been informational.

Structure

The NAPLAN writing task is usually presented on a single sheet full of graphics that are, in most cases, largely irrelevant to the actual assignment. The actual prompt is usually two to four sentences followed by what appear to be instructions for writing a five-paragraph essay followed by bullet points urging students to plan their writing, be careful in word choice, write in sentences and paragraphs, be careful about spelling and punctuation, and check and edit the writing for clarity. There is never any statement providing context or audience. Students are given a total of 40 minutes to plan, write, and edit the essay. The instructions suggest students spend five minutes planning, 30 minutes writing, and five minutes editing for clarity. Although these times are incongruous with the tasks allotted for them, this small amount of time allotted is by no means unusual with mass writing evaluations. However, the lack of sufficient time for prewriting and revision still severely limits any significant evaluation of the entire writing construct.

Marking

By far the most curious and unexplained aspect of the NAPLAN essay is its marking system, with 10 traits and heterogeneous scales. The complete criteria, excerpted from the NAPLAN 2017 Persuasive Writing Marking Guide (Australian Curriculum Assessment and Reporting Authority, 2017a) is displayed in Appendix J.

- Audience 0-6
- Text structure 0-4
- Ideas 0-5
- Persuasive devices 0-4
- Vocabulary 0-5
- Cohesion 0-4
- Paragraphing 0-3
- Sentence structure 0-6
- Punctuation 0-5
- Spelling 0-6.

Multi-trait scoring systems vary from three traits to, at most, seven, with the seventh trait being either a holistic mark or a mark on presentation. I know of no other marking system that employs 10 traits. Moreover, unlike other marking systems that employ consistent scales, either identical or multiples of the base scales, the scales for the NAPLAN essay vary from 0-3 to 0-6. Because there is no public documentation regarding the design of NAPLAN, there is no way of determining how these differing values were determined. The 10 disparate scales also make marking extremely difficult. Markers prefer a single base scale, commonly 4-6 points. They calibrate themselves on that scale, even if in cases such as the UK A-Levels or the ACT Scaling Test their marks reflect multiples of five base-levels. Given that there are roughly one million NAPLAN essays and given the number of markers and time allotted for marking, a very rough estimation would be that, on average, a marker would mark 10 scripts per hour, or one every six minutes (360 seconds). If we estimate that, on average, a marker takes one-and-a-half minutes (90 seconds) to read a script, that leaves 270 seconds for the marker to make 10 decisions or

27 seconds per mark on four different scales. It is inconceivable that markers will consistently make 10 independent decisions in such a short time.

Consequently, the marks will blend into each other. The correlation matrix displayed in Table 3 confirms this hypothesis. With the exception of punctuation, the other traits correlate with each other at 0.7 or greater, producing a shared variance, or overlap of variation, of each of the two variables of 50 per cent or greater (Table 4). Moreover, the mark on the first criterion, audience, appears to have a significant halo effect on subsequent marks, except for punctuation. The variation in scores on each of the other criteria matches approximately two-thirds, 66 per cent or greater, of the variation of audience. Possibly, this first mark stays in the marker's mind and produces a halo effect, influencing all the subsequent rapid-paced marking of that script.

It is difficult to see how totalling these 10 categories with different weights could represent any commonly held consensus of a writing construct. Writing exists to communicate ideas and information. Yet Ideas is given only five marks, while spelling is given six. There is no evidence of a factor analysis of the kind performed by Diederich, French and Carlton (1961) or any other statistically based origin. We have no idea about the procedure or procedures used to create these categories and scales.

Indeed, much more weight is given to spelling, grammar and other mechanics than to communicating meaning. The total number of marks in NAPLAN's marking scheme is 48. Spelling, punctuation, sentence structure, and paragraphing (just forming paragraphs, Text Structure is a separate criterion) comprise 20 of those marks or 41.6 per cent of the total, twice as much as the weight given by the Smarter Balanced and British Columbia tests. Those two tests each have a single category, conventions, which includes punctuation, capitalisation, grammar, usage, and spelling. The British Columbia test also has a separate Style category that includes word choice and sentence structure, but even if these traits are considered as additions to conventions, NAPLAN still gives significantly more weight to mechanics and less weight to meaning than any of the other tests surveyed.

Marking criteria

The current Marking Guides appear to have been developed in 2010 for the Narrative Essay (Australian Curriculum Assessment and Reporting Authority, 2010) and in 2012 for the Persuasive Essay (Australian Curriculum Assessment and Reporting Authority, 2012). Subsequent Marking Guides are identical, containing the same example scripts (Australian Curriculum Assessment and Reporting Authority, 2013, 2017a). Because the Narrative and Persuasive Marking Guides are more similar than different and because the persuasive genre has been used more often, the following discussion will focus, with one exception, on the persuasive.

The Marking Criteria and Supplementary Materials, and Marking Guide Glossary sections are available in the full report.

The exemplar script

In sum, the NAPLAN writing essay, both its overall structure and marking scheme, is paradoxically both overly complex in its marking but simplistic and highly reductive in measuring of any reasonable formulation of the writing construct. Teaching to this test will make students poor writers by having them focus on non-essential tasks such as memorising spelling lists. NAPLAN's influence in the classroom could even negatively affect Australia's standing in international test scores.

The most effective way to support the above assertion is to examine the exemplar script in the Persuasive Writing Marking Guide from at least 2012 to the present. By "exemplar script", I mean a perfect or near perfect script included in the training sample. In the case of the NAPLAN Persuasive Writing Marking Guide , it is a script titled by its author "Is Too Much Money Spent on Toys and Games?" (the interrogative form of the prompt "Too Much Money Is Spent on Toys and Games") but listed in the training scripts as "things should be regulated". A typed version of the script with my commentary emphasising the spelling words appears in Figure 2. The original handwritten copy of the script and the annotations with scores from the Marking Guide are displayed in Appendix L. The script achieved an almost perfect score, 47 out of 48 marks, losing one point for punctuation.

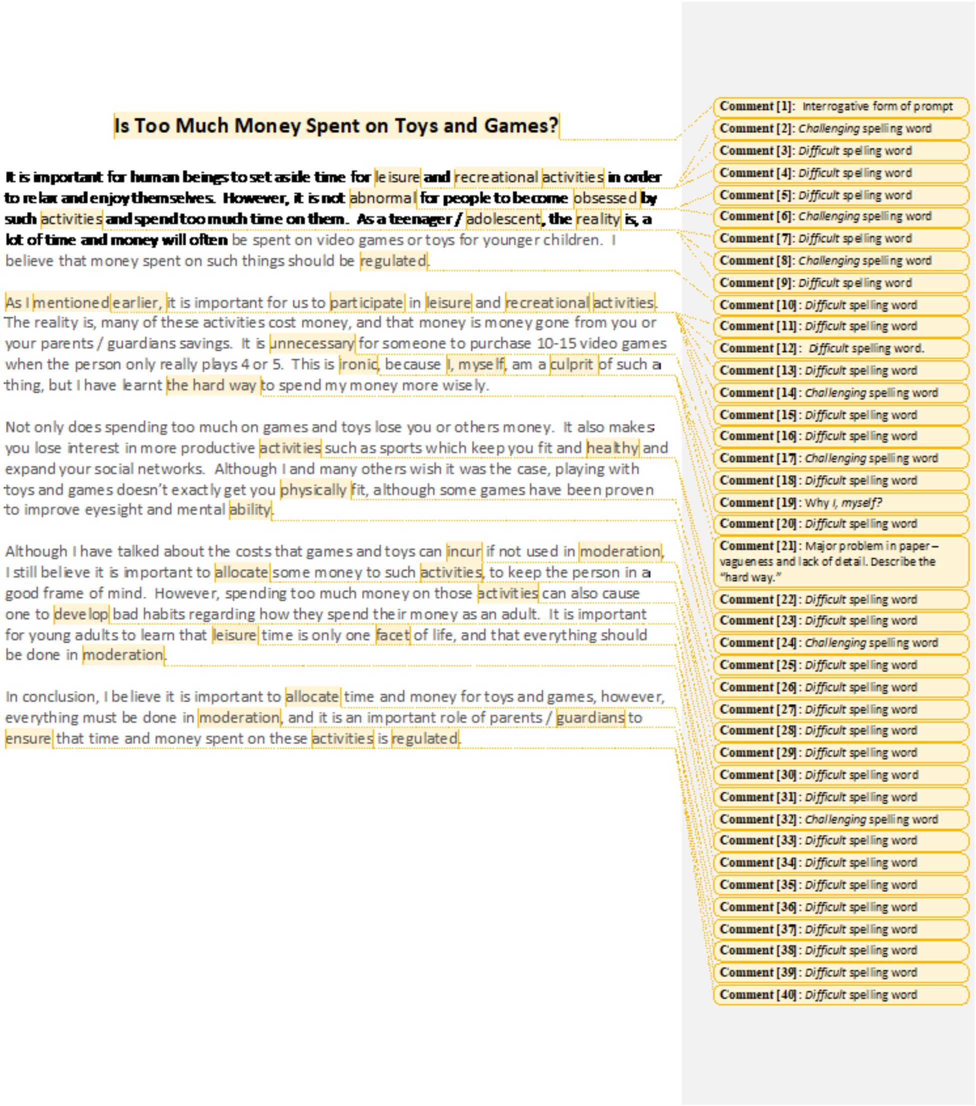
The essay does not say much. Most of the essay could easily be reduced to something like:

> People need to relax and enjoy themselves. Sometimes, however, they spend too much time on such activities. It is unnecessary to purchase ten to fifteen video games when a person only plays four or five. Parents need to control how much time and money their children spend on toys and video games. Although video games may improve eyesight and mental ability, they may detract from playing sports, which promotes fitness and social interaction.

This version is 73 words instead of the 371 words in the original. It is missing a few details. In the original, the author admits that he or she has spent too much money on video games and "learnt the hard way to spend my money more wisely". However, that is all that is said about the incident.

JPL JOURNAL OF PROFESSIONAL LEARNING
www.cpl.asn.au/journal

Click here to access Exemplar persuasive script with Difficult & Challenging spelling words marked



Is Too Much Money Spent on Toys and Games?

It is important for human beings to set aside time for leisure **and** recreational activities **in order to relax and enjoy themselves. However, it is not** abnormal **for people to become** obsessed **by such** activities **and spend too much time on them. As a teenager /** adolescent**, the** reality **is, a lot of time and money will often** be spent on video games or toys for younger children. I believe that money spent on such things should be regulated.

As I mentioned earlier, it is important for us to participate in leisure and recreational activities. The reality is, many of these activities cost money, and that money is money gone from you or your parents / guardians savings. It is unnecessary for someone to purchase 10-15 video games when the person only really plays 4 or 5. This is ironic, because I, myself, am a culprit of such a thing, but I have learnt the hard way to spend my money more wisely.

Not only does spending too much on games and toys lose you or others money. It also makes you lose interest in more productive activities such as sports which keep you fit and healthy and expand your social networks. Although I and many others wish it was the case, playing with toys and games doesn't exactly get you physically fit, although some games have been proven to improve eyesight and mental ability.

Although I have talked about the costs that games and toys can incur if not used in moderation, I still believe it is important to allocate some money to such activities, to keep the person in a good frame of mind. However, spending too much money on those activities can also cause one to develop bad habits regarding how they spend their money as an adult. It is important for young adults to learn that leisure time is only one facet of life, and that everything should be done in moderation.

In conclusion, I believe it is important to allocate time and money for toys and games, however, everything must be done in moderation, and it is an important role of parents / guardians to ensure that time and money spent on these activities is regulated.

Comment [1]: Interrogative form of prompt
Comment [2]: *Challenging* spelling word
Comment [3]: *Difficult* spelling word
Comment [4]: *Difficult* spelling word
Comment [5]: *Difficult* spelling word
Comment [6]: *Challenging* spelling word
Comment [7]: *Difficult* spelling word
Comment [8]: *Challenging* spelling word
Comment [9]: *Difficult* spelling word
Comment [10]: *Difficult* spelling word
Comment [11]: *Difficult* spelling word
Comment [12]: *Difficult* spelling word.
Comment [13]: *Difficult* spelling word
Comment [14]: *Challenging* spelling word
Comment [15]: *Difficult* spelling word
Comment [16]: *Difficult* spelling word
Comment [17]: *Challenging* spelling word
Comment [18]: *Difficult* spelling word
Comment [19]: Why *I, myself?*
Comment [20]: *Difficult* spelling word
Comment [21]: Major problem in paper – vagueness and lack of detail. Describe the "hard way."
Comment [22]: *Difficult* spelling word
Comment [23]: *Difficult* spelling word
Comment [24]: *Challenging* spelling word
Comment [25]: *Difficult* spelling word
Comment [26]: *Difficult* spelling word
Comment [27]: *Difficult* spelling word
Comment [28]: *Difficult* spelling word
Comment [29]: *Difficult* spelling word
Comment [30]: *Difficult* spelling word
Comment [31]: *Difficult* spelling word
Comment [32]: *Challenging* spelling word
Comment [33]: *Difficult* spelling word
Comment [34]: *Difficult* spelling word
Comment [35]: *Difficult* spelling word
Comment [36]: *Difficult* spelling word
Comment [37]: *Difficult* spelling word
Comment [38]: *Difficult* spelling word
Comment [39]: *Difficult* spelling word
Comment [40]: *Difficult* spelling word

Explaining in detail how this person "learnt the hard way" would have produced a much more vivid, memorable, and, probably, more effective essay. What is getting in the way of such development?

The answer is one word, spelling. The annotation on spelling is clear. "Correct spelling of all words. Text meets the requirements for Category 6" (p77). The 20 correct Difficult words and five correct Challenging words are then listed, although a few, such as abnormal, are missed. Figure 2 graphically displays the motive behind the essay's style — pack the essay with as many words categorised as Difficult and Challenging as possible. The doubling of terms such as "teenager/adolescent" and "parents/guardians" is done because only the second term of each pair counts. Redundant use of these words appears to be rewarded. The Difficult spelling word "activities" occurs twice in each of the first two paragraphs and once in each of the following three. The markers are clearly trained to reward such scripts. This script was the only one in the training samples to receive a Category 6 for spelling but the next eight highest scoring scripts all received a Category 5.

What kind of text does this devotion to difficult spelling words produce? I sent this paper to my colleague and mentor, Edward M. White, the person who developed and directed the original holistic scoring of the National Assessment of Educational Progress (NAEP) essays and author of numerous books and articles on writing assessment. Here is his response:

This is a curious paper. It reads as if written by a computer program to meet all the requirements, but it lacks a human voice. It reminds me of Lionel Trilling's definition of basic education: the minimum amount of reading and writing skill for a population to be effectively controlled. If it was written by an actual person, he or she is trained to be a submissive employee. (Email correspondence, 18 January, 2018)

The importance and effect of training sample papers should not be discounted. More than any other element, training sample papers define how markers at any NAPLAN scoring session decide on marks for each script they read.

Defects

Comparison of other Anglophone governmental and non-government organisation essay tests along with an analysis of the NAPLAN essay itself demonstrate that the NAPLAN essay is severely defective in both its design and execution.

- There is a complete lack of transparency in the development of the NAPLAN essay and grading criteria. There is no publicly available document that presents the rationale for the 10 specific criteria used in marking the NAPLAN essay and the assignment of their relative weights. This lack of transparency is also evident in the failure of ACARA to include other stakeholders, teachers, local administrators, parents, professional writers, the business community and others in the formulation, design and evaluation of the essay and its marking criteria.

- Informative writing is not assessed although explicitly included in the writing objectives of the Australian National Curriculum. Informative writing is probably the most common and most important genre both in academic and professional writing. Because that which is tested is that which is taught, not testing informative writing devalues it in the overall curriculum.

- Ten marking criteria with different scales are too many and too confusing, causing high-level attributes such as ideas, argumentation, audience, and development to blend into each other even though they are marked separately. Given the number of markers and time allotted for marking approximately one million scripts, a very rough estimation would be that, on average, a marker would mark 10 scripts per hour, or one every six minutes (360 seconds). If we estimate that, on average, a marker takes one-and-a-half minutes (90 seconds) to read a script, that leaves 270 seconds for the marker to make 10 decisions or 27 seconds per mark on four different scales.

- The weighting of 10 scales appears to be arbitrary. The 10 traits are marked on four different scales, 0-3 to 0-6, and then totalled to compute a composite score. Curiously, the category Ideas is given a maximum of five marks while Spelling is given a maximum of six.

  — There is too much emphasis on spelling, punctuation, paragraphing, and grammar at the expense of higher-order writing issues. While mastery of these skills is important, the essential function of writing is the communication of information and ideas.

  — The calculation of the spelling mark, in particular, may be unique in Anglophone testing. It is as concerned with the presence and correct spelling of limited sets of words defined as Difficult and Challenging as it is with the absence of misspelled words. Markers are given a Spelling reference list categorising approximately 1000 words as Simple, Common, Difficult and Challenging . The scale for the spelling criterion is 0-6. A script containing no conventional spelling scores a 0, with correct spelling of most simple words and some common words yielding a mark of two. To attain a mark of six, a student must spell all words correctly, and include at least 10 difficult words and some challenging words or at least 15 difficult words.

- The NAPLAN grading scheme emphasises and virtually requires the five-paragraph essay form. Although the five-paragraph essay is a useful form for emerging writers, it is extremely restrictive and formulaic. Most arguments do not have three and only three supporting assertions. More mature writers such as those in Year 7 and Year 9 should be encouraged to break out of this form. The only real advantage of requiring the five-paragraph essay form for large-scale testing appears to be that it helps ensure rapid marking.

- Although Audience is a criterion for marking, no audience is defined. There is a significant difference between a generic reader and a specific audience, a distinction that the current NAPLAN essay ignores but is essential for effective writing.

- Specificity in marking rubrics on issues of length and conventions not only skews the test towards low-level skills, it also makes the test developmentally inappropriate for lower years or stages. Several of the marking criteria specify at least one full page as “sustained writing” or “sustained use” necessary for higher marks. It is unrealistic to expect most Year 3 students to produce a full page of prose in 40 minutes.

- The supplementary material provided to markers on argument, text and sentence structure, and other issues is trivial at best and incorrect at worst. It should to be redone entirely as part of the redesign of the NAPLAN essay. Markers should be surveyed to discover what information would be most useful to them.

- The 40 minutes students have to plan, write, revise, and edit precludes any significant planning (prewriting) or revision, two crucial stages of the writing process.

In conclusion, the NAPLAN essay assessment is poorly designed and executed in comparison with similar assessments in Canada, the United States or the United Kingdom. In particular, its focus on low-level skills causes it to de-emphasise the key components of effective written communication. It is reductive and anachronistic. Extending the language of psychometrics, much of the NAPLAN essay marking is not only construct irrelevant, some of its features, such as rewarding the use of a narrowly defined list of words, are construct antagonistic.

Dr Perelman's guide to a top-scoring NAPLAN essay

More than 10 years ago, I was a vocal critic of the SAT writing test. At the behest of some MIT students who were tutoring inner-city high school students for the test, I developed Dr Perelman's SAT Essay Writing Tips to help the high school students perform well on the test but also to emphasise to them that the writing that would receive a high mark was formulaic, artificial, and had little to do with real-world or even real academic writing. The students did well, and the tips went viral over the web (Perelman, nd). Unexpectedly, they became a powerful tool in my efforts to end what I considered a test that subverted, not supported, instruction in effective writing. Students followed my instructions and received high scores in the essay, knowing full well that they were not writing "real" essays but just gaming the system, transforming, for some, an awe about the test into intense cynicism.

The tips were noticed by David Coleman, the incoming President of the College Board, and he invited me to come down to New York and talk about the problems inherent in the test essay and the possibility of a new writing exercise. As a consequence, the old SAT essay was abolished. What this experience taught me most was that by publicly showing students how easy it is to game such tests, I was extremely effective in exposing the shortcomings of and the contradictions within these exercises.

My study of the NAPLAN essay marking has produced a similar conclusion about the disassociation of the NAPLAN marking scheme from any authentic construct of writing ability. Moreover, its emphasis on form and the correct spelling of certain words makes it even easier to provide students with construct irrelevant strategies to attain high marks. There are three reasons for releasing Dr Perelman's guide to a top-scoring NAPLAN essay. First, I am sure that such strategies already exist in some classrooms. That which is tested always informs that which is taught. Making them public democratises opportunity on NAPLAN. Second, such advice exposes the poor pedagogical practices that are encouraged by the test. Simultaneously, when students use them and score well, it reveals which constructs are being assessed and which constructs are not. The one-page Dr Perelman's guide to a top-scoring NAPLAN essay appears in Figure 3 and the spelling reference list appears at the end of Appendix K.

The guide pertains to persuasive essays with one exception. Because the Narrative Marking Guide instructs readers to ignore "derivative texts", (ie scripts that appropriate a plot from a book, film, or TV program) "the student's work must be marked on its merits as an original script" (Australian Curriculum Assessment and Reporting Authority, 2010, p72). The explanation for this policy is that not every reader would know the original texts. So, for the sake of consistency, markers are told to ignore their recognition. Of course, the same argument could be made (and often is) in cases of academic plagiarism. Because some teachers have graded narrative scripts and know of this policy, some may have informed their students. In addition, the policy is in the Marking Guide available on the web. By mentioning this rule, I am only trying to level the playing field.

Dr Perelman's guide to a top scoring NAPLAN essay

1. Memorise the list of Difficult and Challenging Spelling Words and sprinkle them throughout the paper. Feel free to repeat them, and do not worry very much about the meaning.
2. If you are not sure how to spell a word, do not use it.
3. Repeat the language and ideas in the Writing Task throughout the paper.
4. Begin at least one sentence with the structure, "Although x (sentence), y (sentence)." For example: "Although these instructions are stupid, they will produce a high mark on the NAPLAN essay."
5. Master the five-paragraph form.

    a) Have a minimum of four paragraphs, preferably five.
    b) Each paragraph, except the last one, should have a minimum of four sentences. Do not worry about repeating ideas.
    c) The first paragraph should end with your thesis sentence.
    d) The next-to-last paragraph should modify your thesis sentence by taking the other side of the issue in special cases.
    e) The last paragraph should begin with "In conclusion" and then repeat the thesis sentence from the first paragraph. Then just repeat two or three ideas from the other paragraphs.

6. Increase your score on the "Audience" and "Persuasive Devices" categories by addressing the reader using "you" and ask questions. For example: "So you think you wouldn't mind writing a stupid essay?"
7. Use connective (Velcro) words such as "Moreover," "However," "In addition", "On the other hand" at the beginning of sentences.
8. Begin sentences with phrases such as "In my opinion", "I believe that", "I think that" etc.
9. Repeat words and phrases throughout your paper.
10. Employ the passive voice frequently throughout your paper.

11. Use referential pronouns, such as "this", without a reference noun following it. For example, "This will make the marker think you are a coherent writer".
12. Make arguments using forms such as "We all believe that we should do X" or "We all know that Y is harmful".
13. Always have at least one, preferably two adjectives next to nouns. Thus, not "the dog" but the "frisky and playful dog".
14. If you are writing a narrative essay, think quickly if there is a television program, movie, or story that you know that fits the requirements of the narrative writing task. If there is one use it as your narrative, embellishing it or changing it as much as you want. Markers are explicitly instructed to ignore if they recognise any stories or plots and mark the script on its own merits as if it was original.
15. Never write like this except for essay tests like the NAPLAN.

Developing a new NAPLAN writing assessment

NAPLAN should not be discarded but reformulated and reimagined to promote and reinforce the curriculum and classroom teaching. If all three are aligned, then teaching to the test ceases to be a problem and becomes the way things should be. My expertise is in writing assessment. Consequently, my discussion focuses only on the development of a new NAPLAN essay, although some of my discussion will be relevant to the development of other components. The following discussion is divided into two parts. The first, an outline of a process for development, is a tentative recommendation. The second, the description of one possible implementation, is given solely as an example or as a vision to help begin discussion.

The Developing a New NAPLAN Writing Assessment section is available in the full report.

Australia produces great language assessments. I admire the various Australian state and territory English and writing HSC papers. IELTS, developed in Australia and the United Kingdom, is by far the best test of English as a foreign language. Australia can produce a great NAPLAN essay assessment.

The Work Cited section is available in the full report.

Les Perelman is an internationally recognised expert in writing assessment and the application of technologies to assess writing. He has written opinion pieces for The Boston Globe, The Washington Post and The Los Angeles Times. He has been quoted in The New York Times, The New Yorker, The Chicago Tribune, The Boston Globe, The Los Angeles Times and other newspapers. Dr Perelman has been interviewed on television by ABC, MSNBC and NHK Japan Public Television, and interviewed on radio by National Public Radio, various NPR local stations, the Canadian Broadcasting Corporation, and the Australian Broadcasting Corporation.

The President of the College Board has credited Dr Perelman's research as a major factor in his decision to remove and replace the writing section of the SAT. Dr Perelman is a well-known critic of Automated Essay Scoring (AES). To demonstrate the inability of robo-graders to differentiate writing from gibberish, he and three undergraduates developed the "BABEL Generator" , which produces verbose and pretentious nonsense that consistently receives high marks from AES machines.

Dr Perelman received his BA in English Language and Literature from the University of California, Berkeley, and his MA and PhD in English from the University of Massachusetts. After a three-year post-doctoral fellowship in Rhetoric and Linguistics at the University of Southern California, Dr Perelman moved to Tulane University where he served as an Assistant Professor of Rhetoric, Linguistics and Writing, Director of First Year Writing, Director of the Writing Centre and a member of the Graduate Faculty.

For the next 25 years Dr Perelman was Director of Writing Across the Curriculum in Comparative Media Studies/Writing at the Massachusetts Institute of Technology and served as an Associate Dean in the Office of the Dean of Undergraduate Education. He was Project Director and co-principal Investigator for a grant to MIT from the National Science Foundation to develop a model CommunicationIntensive Undergraduate Program in Science and Engineering. He served as principal Investigator for the development of the iMOAT Online Assessment Tool funded by the MIT/Microsoft iCampus Alliance. Dr Perelman has served as a member of the Executive Committee of the Conference on College Composition and Communication, the post-secondary organisation of the National Council of Teachers of English and co-chaired the Committee on the Assessment of Writing. He is currently a member of the editorial board of Assessing Writing.

Dr Perelman has been a consultant to more than 20 colleges and universities on the assessment of writing, program evaluation, and writing across the curriculum. Dr Perelman has served as a consultant for writing program assessment and development for the Fund for the Improvement of Postsecondary Education of the US Department of Education and for the Modern Language Association. In 201213, he served as a consultant to Harvard College and as co-principal investigator in a two-year study assessing the writing abilities of undergraduates at the college.

Dr Perelman co-edited the volume Writing Assessment in the 21st Century and he is the primary author of the first web-based technical writing handbook, The Mayfield Handbook of Technical and Scientific Writing. He has published articles on writing assessment, technical communication, computers and writing, the history of rhetoric, sociolinguistic theory and medieval literature, and co-edited The Middle English Letter of Alexander to Aristotle.