TOWARDS A NEW NAPLAN: TESTING TO THE TEACHING

Les Perelman, Ph.D.

Towards a new NAPLAN: Testing to the Teaching
By Les Perelman, Ph.D.

Published 2018

ISBN 978-0-3482555-2-9

Authorised by John Dixon, General Secretary, NSW Teachers Federation, 23-33 Mary Street, Surry Hills 2010. 18116.

Cover photo: istock.com/Steve Debenport

TOWARDS A NEW NAPLAN: TESTING TO THE TEACHING

Les Perelman, Ph.D.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Click here for appendices

EXECUTIVE SUMMARY

Achievement tests have become an almost universal feature of primary and secondary education in industrialised countries. Such assessments, however, always need to be periodically reassessed to examine whether they are measuring the relevant abilities and whether the results of the assessment are being used appropriately. Most importantly, the assessments must themselves be assessed to ensure they are supporting the targeted educational objectives. Contemporary concepts of validity are considered as simultaneous arguments involving the interpretation of construct validity, content validity, and external validity, along with arguments involving fairness and appropriateness of use.

As points of comparison, the examination of six different writing tests from the United States, Australia, Canada and the United Kingdom produced observations directly relevant to an evaluation of the NAPLAN essay:

- **The majority of tests, and all the tests specifically for primary and secondary schools, are developed, administered, and refined within the context of publicly available framework and specification documents.** These documents articulate, often in great detail, the specific educational constructs being assessed and exactly how they will be measured. They are not only an essential tool for any assessment design but also their publication is vital for the transparency and accountability necessary for any testing organisation.
- **In some cases, these documents are produced with collective input from stakeholders and academic specialists in the specific disciplines.** The Smarter Balanced Assessment Consortium and the National Assessment of Educational Progress (NAEP) writing assessments made use of large panels of teachers, administrators, parents, elected officials, and academic experts.
- **Several of the tests unreservedly mix reading and writing.** The Smarter Balanced Assessment Consortium reading test incorporates short-answer writing (constructed response). The texts in the reading exercise form part of the prompt for the long essay, and the short written answers to the reading questions serve as prewriting exercises. Integrating writing and reading in assessments makes sense. Children acquire language through exposure to speech. Eventually, reception leads to production. Although writing is a technology that is only approximately 6000 years old, it is an analogue to speech, albeit not a perfect one. Indeed, students will have extreme difficulty writing in a genre if they have not read pieces in that same genre.
- **Writing tasks are designed and employed for specific classes or years.** With the exception of NAPLAN, I know of no other large-scale writing assessment that attempts to employ a single prompt for different age groups.
- **Similarly, most tests tailor their marking rubrics for different classes or years.** For example, the scoring rubrics for Grades 4 and 7 in British Columbia's Foundation Skills Assessment (FSA), displayed in Appendix D (see online report), vary significantly, consistently expecting a higher level of performance from the higher grade.
- **Informative writing, in addition to narrative and persuasive writing, is a common genre in school writing assessments.** Much of the writing students will do in school and then in higher education and in the workforce will be informative writing.
- **Several of the assessments explicitly define an audience and, often, a genre as part of the writing task.** One prompt from the National Assessment of Educational Progress (NAEP) assessments asks students to write a letter to the school principal on a specific issue. A Smarter Balanced Assessment Consortium informative writing task for Grade 6 students asks the student to write an informative article on sleep and naps (the topics of the reading questions) for the school newspaper that will be read by parents, teachers, and other students.

- **All of the other assessments that employ multi-trait scoring use the same or similar scales for all traits. Moreover, they all employ significantly fewer trait categories.** The Smarter Balanced Assessment Consortium employs three scales: two are 1-4, and the Conventions scale is 0-2. British Columbia's Foundation Skills Assessment uses five scales, all 1-4. The Scholastic Aptitude Test (SAT) has three 1-4 scales that are not summed, and UK tests such as A and AS Levels have multiple traits, usually four to six, that are always scored on scales that are multiples of 1-5 levels.
- **Most of the assessments, and all of the assessments that focused on the primary and secondary years/grades, allowed students access to dictionaries and, in some cases, grammar checkers or thesauri.** Some of the assessments are now on networked computers or tablets that include standard word processing applications with spell-checkers or dictionaries and other tools for writing.

Comparison of other Anglophone governmental and non-government organisation essay tests along with an analysis of the NAPLAN essay demonstrate that the NAPLAN essay is defective in 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 the Australian Curriculum Assessment and Reporting Authority (ACARA) to include other stakeholders, such as teachers, local administrators, parents, professional writers, 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 Curriculum.** Informative writing is probably the most common and most important genre in both 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. It is inconceivable that markers will consistently and accurately make 10 independent decisions in such a short time.
- **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 5 marks while Spelling is given a maximum of 6.
  - **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 2. To attain a mark of 6, 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 to ensure rapid marking.
- **Although "audience" is a criterion for marking, no audience is defined in the writing prompt.** 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 summary, the NAPLAN essay fails to be a valid measure of any serious formulation of writing ability, especially within the context of its current uses. Indeed, NAPLAN's focus on low-level mechanical skills, trivialisation of thought, and its overall disjunction from authentic constructs of writing may be partially responsible for declining scores in international tests.

There should be an impartial review of NAPLAN, commencing with special attention being paid to the writing essay, leading to a fundamental redesign of the essay and the reconsideration of its uses. Such a review should also consider the way in which NAPLAN is administered and marked, its current disconnection to a rich curriculum and the specific and diverse teaching programs that children experience in classrooms.

Such a review should be an inclusive process encompassing all elements of the educational and academic communities with the key focus areas identifying the particular needs of students, how they have progressed in their class with the programs they are experiencing and how systems, jurisdictions and the nation can further support their intellectual growth and futures. A principal emphasis in this review should be to promote alignment of the curriculum, classroom pedagogy, and all forms of assessment; that is, to test to the teaching. If students consider classroom exercises and outside assessments to be indistinguishable, and both reflect the curriculum, then assessments reinforce teaching and learning rather than possibly subverting them.

Australia produces great language assessments. I admire the various Australian state and territory English and writing HSC papers. The International English Language Testing System (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 major writing 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 large-scale 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.

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 refutes 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 task-based 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: NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (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 one-sided 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.

National Assessment of Educational Progress

Like the international PISA tests, the National Assessment of Educational Progress (NAEP) is designed as a representative periodic assessment of student abilities in various subjects, including writing. Since 1969, the NAEP has been assessing the writing abilities of Grade 4, 8 and 12 students every four to six years (National Center for Educational Statistics, 2017). Computer-based writing was instituted in 2011 for Grades 8 and 12 and piloted in 2012 for Grade 4. In composing on computers, students have the full use of the computer-based tools for writing, including spelling and grammar checkers. The 2011 assessment included nationally representative samples of 24,100 Grade 8 students and 28,100 Grade 12 students (National Center for Educational Statistics, 2017). In 2017, writing assessments were administered to similar sample populations of Grade 4 and Grade 8 students. As displayed previously in Table 1, each group writes two essays, each in response to a different writing task representing one of the three purposes of writing as defined by the NAEP:

- to persuade, in order to change the reader's point of view or affect the reader's action
- to explain, in order to expand the reader's understanding
- to convey experience, real or imagined, in order to communicate individual and imagined experience to others (National Assessment Governing Board, 2010).

Sample writing tasks are included in Appendix A (see online report). These three types of tasks are, of course, close approximations of the traditional three genres of school writing, persuasive, informative, and narrative. They differ, however, in that the language emphasises writing as a communicative action directed to a specific reader or readers. Indeed, the National Assessment of Educational Progress (NAEP) 2011 Frameworks mandate that the "writing task specify or clearly imply an audience" (National Assessment Governing Board, 2010 p.vi). A chart outlining the overall content components of the assessment is displayed in Appendix B (see online report).

In the 2011 assessment, there were a total of 22 separate writing tasks for Grade 8 and 22 for Grade 12. Students were randomly assigned two different types of writing tasks and given 30 minutes to complete each of them (National Center for Education Statistics, 2011). The random assignment and combinations of tasks among the three genres greatly reduced the potential for prompt bias for specific subgroups of students, a significant problem when only one prompt is used but possibly appropriate for an assessment that is not being used to rank individual students or schools. The random assignment of two of the three genres also ensures that students will write at least one of the two expository genres, to persuade or to explain. Moreover, the variation in prompts can provide valuable data on the effect of specific features on various groups and subgroups that can be used to inform future assessments.

TABLE 2: COMPARATIVE ANGLOPHONE SCHOOL WRITING ASSESSMENTS

Test	Country	Use	Score type	# Markers	Who marks	Year or Grade	Time (mins)	Computer	Other writing	Preparatory writing	Type of prompt	Dictionary	Thesaurus	% Conventions
NAPLAN	Australia	Snapshot; school evaluation; school choice; student feedback	Multi-trait 10 composite; different scales	1	Teachers mass marking	3, 5, 7, 9	40	No?	No	No	Short general	No	No	42%
NAEP	USA	Snapshot	Holistic	2	Teachers and others mass marking	4, 8, 12	2 tasks, 30 mins each	Yes	Yes	No	Task with audience	Yes (online)	No	Holistic
Smarter Balanced	USA	School evaluation; student evaluation	Multi-trait composite weighted; component of 3 scores	1	Teachers mark online	3-8, 11	120	Yes	Yes	Yes	Genre task (e.g. newspaper article)	Yes (online)	Yes (online)	Approximately 20%
Foundation Skills Assessment — British Columbia	Canada	School evaluation (procedures being reconsidered); student feedback	Composite of holistic + 4 traits	1	Local teachers	4,7	45 Students given more time if needed	No	30 min short answer questions and reflective essay at end	No	Short general	Yes	Yes	20%
SAT	USA	University admission	Multi-trait 3 reported separately	2	Online graders	12	50	No	No	No	Rhetorical and argumentative analysis	No	No	Writing score includes Conventions
English Language A-Levels AQA	UK	University admission	Multi-trait marks	1	Teachers mark online	12	150	No	Yes	No	Writing in a specific genre and linguistic register	No	No	Writing score including Conventions is 10% of Paper 2
ACT Scaling Test	Australia	University admission	Holistic 20-point scale at 5 levels like UK	4	Teachers	12	150	No	No	No	Open-ended stimulus of several pages	Yes	No	Holistic

Les Perelman, PH.D. 22 January 2018

Smarter Balanced Assessment Consortium

The Smarter Balanced Assessment Consortium was created in 2010 as one of two US state consortia to assess the Common Core Standards. Currently, 13 states, the Bureau of Indian Education, and the United States Virgin Islands are members of the Consortium (Smarter Balanced Assessment Consortium, 2017). The writing tasks in the Smarter Balanced Assessment Consortium assessment attempt to evaluate a range of elements that comprise the writing construct, including invention (prewriting); the ability to synthesise, evaluate, and use information as evidence; rhetorical awareness of specific audiences and genres; revision; and adherence to conventions (Smarter Balanced Assessment Consortium, 2015). The Smarter Balanced Assessment Consortium assessments are administered on computers, and students write the essays in a word processing application containing an online spell-checker and thesaurus.

These very comprehensive goals are made possible by an approach that seamlessly unites the two elements of literacy, reading and writing. Students are required to read several sources on a topic, complete a series of short answer questions, and then complete a writing task that involves synthesising and using the information from the readings to write in a specific genre for a defined audience. One set of exercises for Grade 3 students first contains several sources about astronauts, followed by this prompt:

> Your teacher is creating a bulletin board display in the school library to show what your class has learned about different types of jobs. You decide to write an informational article on astronauts. Your article will be read by other students, teachers, and parents.
>
> Using more than one source, develop a main idea about being an astronaut. Choose the most important information from the sources to support your main idea. Then, write an informational article that is several paragraphs long. Clearly organise your article and support your main idea with details from the sources. Use your own words except when quoting directly from the sources. Be sure to give the source title or number when using details from the sources (Smarter Balanced Assessment Consortium, nd, *Item Key* 2558).

A sample set of exercises for Grade 6 students, contains informational readings about hiking the Grand Canyon. The extended essay prompt asks for an imaginative narrative of an exciting adventure hiking the Grand Canyon to be published in an online magazine that will be read by parents, teachers, and students. Students are instructed to use information and details from the sources in writing their story (Smarter Balanced Assessment Consortium, nd, *Item Key* 3678). A Grade 8 argumentative question asks students, after they have read several articles about the US penny (1 cent) coin, to write an argumentative essay either for or against its continued use (Smarter Balanced Assessment Consortium, nd, *Item Key* 2698). Students have two hours to read the short sources, respond to the short answer questions, and write the essay. Because the short responses also serve as a prewriting activity, the two hours is a minimally reasonable amount of time for students to plan, write, revise and edit an essay.

The essays are graded on three trait scales: Organisation and Purpose, 1-4; Evidence/Elaboration (for informative and persuasive essays), Development/Elaboration (for narrative), 1-4; and Conventions, 0-2. Sample rubrics and grade-based convention charts based on the United States' Common Core State Standards are displayed in Appendix C (see online report). Conventions are marked as a group and comprise approximately 20 per cent of the total score.

Foundation Skills Assessment — British Columbia

The Canadian Province of British Columbia's Ministry of Education employs the Foundation Skills Assessment (FSA) to assess Grade 4 and 7 students' numeracy, reading, and writing skills (British Columbia Ministry of Education, 2017a). The test is used to provide data on student achievement in literacy and numeracy to the province, districts, and schools. The test contains online and handwritten components, with the writing and most of the reading assessments completed on paper. In the Grade 4 test, the writing components include three short-answer responses to two readings on a common theme, one informational and one literary. The third question asks students to consider a specific connection between the two texts. The extended writing essay employs a simple prompt of two or three sentences. Following the prompt, there is a planning page that asks specific short-answer questions to aid in prewriting activities that will help inform the paper. Students can use a dictionary and a thesaurus on any part of the test (British Columbia Ministry of Education, 2016).

Recently, the test has been redesigned (British Columbia Ministry of Education, 2017d). Students have a choice of two sets of readings, each on a specific theme. Before students begin the reading, there is a collaborative classroom activity that involves students in pairs exploring the two themes through visiting posters containing titles and pictures of each of the four readings. Students are asked to converse with their partner about possible connections that can be made between each of the images and titles and what kind of information or ideas might be found in the texts. This activity is followed by a discussion by the entire class (British Columbia Ministry of Education, 2017b). In addition, after both the written segment and the computer-based segment of the entire test, there is a short, non-graded self-reflection activity that involves writing. The structure of the new test is displayed in Figure 1.

Allowing a choice of readings and prompts violates the received wisdom of many testing organisations. The argument is that student choice creates an uneven playing field. However, the growing diversity of testing populations makes a powerful case that choice can improve fairness rather than impede it by allowing individuals to select the readings and prompts that best fit their knowledge base, interests, and learning styles.

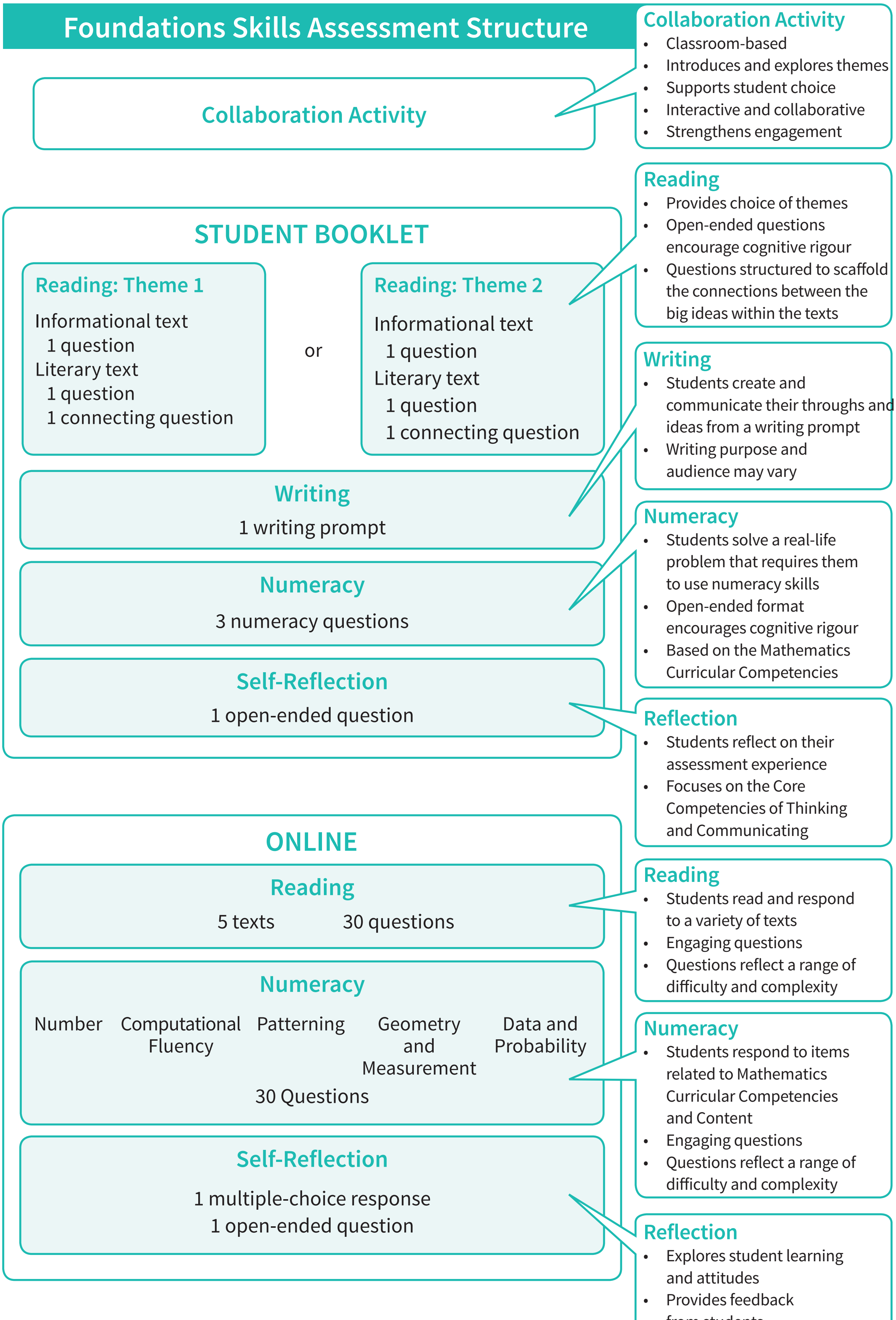
The scripts are given five scores on 1-4 scales: a holistic score (called a snapshot) and trait scores for meaning, style, form, and conventions. The grade 4 and grade 7 rubrics vary slightly in trait definition but are largely similar. The five scores are then summed to produce a composite score. The two rubrics are displayed in Appendix D (see online report).

Unlike most other large state examinations, the Foundations Skills Asssessment is locally marked and the method of marking is locally determined. Some local groups mark individually, others read in pairs, and in other schools marking is done in groups (British Columbia Ministry of Education, 2017c). Each year, local writing scores are closely monitored and monitoring reports are published. These reports substantiate the reliability of this system. In 2013, more than 98 per cent of the local scores on all the writing segments were within 1 point of the score assigned at the monitoring session (British Columbia Ministry of Education, 2013).

SAT essay

The SAT (Scholastic Aptitude Test) essay, like the two that follow it here, differ significantly from the previous three tests in purpose and population. The previous tests assessed student performance at various points during students' primary and secondary education. The SAT and the following two tests are leaving examinations, assessing students' abilities at the end of secondary schooling, often as a major component for entrance to tertiary education. They are included because they exhibit features that are of interest in evaluating the efficacy and reliability of the NAPLAN essay.

FIGURE 1: STRUCTURE OF NEW FOUNDATIONS SKILLS ASSESSMENT (BRITISH COLUMBIA MINISTRY OF EDUCATION, 2017a)



Foundations Skills Assessment Structure

Collaboration Activity

Collaboration Activity
- Classroom-based
- Introduces and explores themes
- Supports student choice
- Interactive and collaborative
- Strengthens engagement

STUDENT BOOKLET

Reading: Theme 1

Informational text
1 question
Literary text
1 question
1 connecting question

or

Reading: Theme 2

Informational text
1 question
Literary text
1 question
1 connecting question

Reading
- Provides choice of themes
- Open-ended questions encourage cognitive rigour
- Questions structured to scaffold the connections between the big ideas within the texts

Writing

1 writing prompt

Writing
- Students create and communicate their throughs and ideas from a writing prompt
- Writing purpose and audience may vary

Numeracy

3 numeracy questions

Numeracy
- Students solve a real-life problem that requires them to use numeracy skills
- Open-ended format encourages cognitive rigour
- Based on the Mathematics Curricular Competencies

Self-Reflection

1 open-ended question

Reflection
- Students reflect on their assessment experience
- Focuses on the Core Competencies of Thinking and Communicating

ONLINE

Reading

5 texts 30 questions

Reading
- Students read and respond to a variety of texts
- Engaging questions
- Questions reflect a range of difficulty and complexity

Numeracy

Number | Computational Fluency | Patterning | Geometry and Measurement | Data and Probability

30 Questions

Numeracy
- Students respond to items related to Mathematics Curricular Competencies and Content
- Engaging questions
- Questions reflect a range of difficulty and complexity

Self-Reflection

1 multiple-choice response
1 open-ended question

Reflection
- Explores student learning and attitudes
- Provides feedback from students

Foundation Skills Assessment: Description and Specifications — June 2017

In 2005, the College Board introduced a 25-minute essay as part of a new SAT writing section. There was immediate opposition to the change, largely based on criticisms that essays written in 25 minutes on short general prompts would not address much of any accepted formulation of the writing construct, but instead would reward essays largely on length alone (Perelman, 2012; Winerip, 2005). In 2014, partially in response to these criticisms, David Coleman, the President of the College Board, announced that the SAT essay would be discontinued and a new, optional SAT essay based on a reading would be offered beginning in 2016 (Balf, 2014). [Full Disclosure: I was one of the major critics of the 2005 SAT essay section and have been credited by Mr Coleman as influential in his decision to discard it. I also informally participated in discussions involving the design of the new SAT essay.]

The new SAT essay is optional. Rather than the 25 minutes of the old mandatory essay, students have 50 minutes. The writing task is always based on a short, published opinion piece that makes a definite argument. Unlike many prompts — such as those used for the persuasive essays in NAPLAN that ask a writer to agree, disagree, or, in some tests, both agree and disagree with the author's argument — the new SAT essay asks for an analysis of the argument rather than an argument in response. Specifically, the prompt always asks students how the author employs logic, rhetoric and stylistic devices to persuade the reader (College Board, 2015, 2018).

This assessment clearly asserts the interconnection between reading and writing. The essay is evaluated on these traits: reading, analysis and writing. Each trait is scored by two readers on 1-4 scales that are displayed in Appendix E (see online report). Rather than producing a single composite score, the readers' scores on each trait are combined to produce three trait scores on a 2-8 scale (College Board, 2017). The emphasis is on understanding and conveying how meaning and arguments are conveyed through the written word.

United Kingdom English Language A Levels (AQA)

In the United Kingdom, there are several examination boards — AQA, EDEXCEL, OCR, EDUQAS, WJEC — each recruiting its own markers and each with different kinds of training overseen by the Office of Qualifications and Examinations Regulation (OFqual). These examination scores serve as an important component in university admission. Most of the marking is now done online by teachers and university lecturers. Students write several scripts in what constitutes a single "paper". Writing ability was one of the traits assessed in the 2014 Language A-Level Paper 2. The marking scheme displayed in Appendix F (see online report) contains the basic structure of all A and AS level exams. Each script may receive marks for several rubrics. The question displayed in this marking scheme addresses two assessment objectives:

- demonstrate critical understanding of concepts and issues relevant to language use
- demonstrate expertise and creativity in the use of English to communicate in different ways. (AQA, 2014)

The marking grid displayed in Appendix F (see online report) exhibits common characteristics of United Kingdom A and AS level marking. The basic metric consists of five levels. Differences in marking schemes are done in multiples of the five levels, in this case 10 marks, which consist of a high and low mark at each level. Markers are always instructed to determine level first and then make a finer determination by selecting the appropriate mark.

One major component of the score is genre knowledge and adherence to the specified genre. In the specific case displayed in Appendix F (see online report), the genre is an opinion piece, not a five-paragraph academic essay. Indeed, the marking scheme directs the reader to give lower marks to responses that rely on paragraphing or academic language.

ACT Scaling Test

The ACT is the tertiary test given in the Australian Capital Territory. The test booklet given to students is displayed in Appendix G (see online report) and the Marker Briefing in Appendix H (see online report). The writing test is marked on a 20-point scale. However, the base of this scale is a five-level scale, E, D, C, B, A with gradations of three marks within each level. This test has several remarkable features. First, there is no specific prompt or question. Students are presented with two or three pages of stimulus material on a specific issue. In 2017, the issue was trust; previously, topics such as social media were used. The general prompt is:

> Read carefully the material on these two pages. Write about 600 words, giving your point of view on the major issue raised in the material. You need not refer to any of the material specifically, but you must deal with the major issue in it. Do not summarise the material.

Students have to make an argument. The general topic is defined. Various short arguments and data are presented. The student, much like students in university classes or writers in the real world, is not given a ready-made thesis; he or she has to discover it. Thus, the test is able to assess discovery and invention, two essential components of both critical thinking and the writing construct. Second, students are given two-and-a-half hours to write the test, giving enough time for engagement in all stages of the writing process: invention, writing, revision, and editing. Third, rather than having an undefined length, the test gives a 600-word maximum. Limiting the number of words in the essay as well as giving students ample time for revising significantly reduces the importance of length as a determiner of score. Moreover, the open-ended nature of the writing task makes the test less amenable to machine marking. And, indeed, automatic essay scoring machines have been unable to approximate human scores on this test (McCurry, 2010). Finally, the test has the unusual procedure of four markers. Although expensive, the presence of so many markers in totalling the score substantially minimises the effect of one aberrant reader.

Noteworthy features of these tests

A major reason for undertaking this by-no-means exhaustive or representative survey was to examine elements of these tests as alternatives or best practices in comparison to NAPLAN.

- **The majority of tests, and all the tests specifically for primary and secondary schools, are developed, administered, and refined within the context of publicly available framework and specification documents.** The motivations and justification for the redesign of the 2011 National Assessment of Educational Progress (NAEP) Writing Test are carefully articulated in the *Writing Framework for the 2011 National Assessment of Educational Progress.* In 2015, the Smarter Balanced Assessment Consortium published *Content Specifications for the Summative Assessment of the Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects.* As part of the redesign of the Foundation Skills Assessment, the British Columbia Ministry of Education issued the *Foundation Skills Assessment Description and Specifications.* Finally, a major step in the development of the new SAT essay was the production of the document *Test Specifications for the Redesigned SAT®*. These documents articulate, often in great detail, the specific educational constructs being assessed and exactly how they will be measured. They are not only an essential tool for any assessment design; their publication is essential for the transparency and accountability necessary for any testing organisation.
- **In some cases, these documents are produced with collective input from stakeholders and academic specialists in the specific disciplines.** Both the Smarter Balanced Assessment Consortium and NAEP writing assessments made use of large panels of teachers, administrators, parents, elected officials and academic experts.

- **Several of these tests unreservedly mix reading and writing.** Psychometricians instinctively want to isolate constructs so each can be measured without "contamination" by another construct. This position is now being challenged. Smarter Balanced Assessment Consortium incorporates short-answer writing (constructed response) into the reading test. The texts in the reading exercise form part of the prompt for the long essay, and the short written answers to the reading questions serve as prewriting exercises. The Foundation Skills Assessment *Description and Specifications* states*:*

> Reading and writing are interconnected; reading influences writing and writing influences reading. Research has shown that when children read extensively they become better writers. In the process of writing their own texts, students come to better understand an author's construction of his or her texts. This understanding is considered in defining reading and writing. (British Columbia Ministry of Education, 2017a, p4)

The new optional SAT Essay has also been designed with the explicit intention of making the essay not only a writing exercise but also an exercise in reading and textual analysis.

> In a break from the past and present of much standardised direct writing assessment, the essay task is not designed to elicit students' subjective opinions but rather to assess whether students are able to comprehend an appropriately challenging source text and to craft an effective written analysis of that text. Rather than merely asking students to emulate the form of evidence use by drawing on, say, their own experiences or imaginations, the Essay requires students to make purposeful, substantive use of textual evidence in a way that can be evaluated objectively. The Essay also connects reading and writing in a manner that both embodies and reinforces the interdependency of these ELA/literacy skills. (College Board, 2015, p71)

Integrating writing and reading in assessments makes sense. Children acquire language through exposure to speech. Eventually, reception leads to production. Although writing is a technology that is only approximately 6000 years old, it is a close analogue, albeit not a perfect one. Students will have extreme difficulty writing in a genre if they have not read pieces in that same genre. Moreover, as implied by the College Board statement, without providing students with texts or other forms of information, we invite them to support their arguments with made up "alternative facts".

- **Writing tasks are designed and employed for specific classes or years.** With the exception of NAPLAN, I know of no other large-scale writing assessment that attempts to employ a single prompt for different age groups. This practice produced some major issues in 2014 (A. Smith, 2014), which led to subsequent administrations of NAPLAN having different prompts for Years 3 and 5 and for Years 7 and 8.
- **Similarly, most tests tailor their marking rubrics for different classes or years.** For example, the scoring rubrics for Grades 4 and 7 in British Columbia's Foundation Skills Assessment displayed in Appendix D (see online report) vary significantly, consistently expecting a higher level of performance from students in the higher grade.
- **Informative writing, in addition to narrative and persuasive writing, is a common genre in school writing assessments.** Much of the writing students will do in school and then in higher education and in the workforce will be informative writing.
- **Several of the assessments explicitly define an audience and, often, a genre as part of the writing task.** One prompt from the NAEP assessments asks students to write a letter to the school principal on a specific issue. A Smarter Balanced Assessment Consortium explanatory writing task for Grade 6 students asks the student to write an explanatory article on sleep and naps (the topics of the reading questions) for the school newspaper that will be read by parents, teachers, and other students.

- **All of the other assessments that employ multi-trait scoring use the same or similar scales for all traits. Moreover, they all employed significantly fewer trait categories.** Smarter Balanced Assessment Consortium assessments employs three scales: two are 1-4, and the conventions scale is 0-2. British Columbia's Foundation Skills Assessment uses five scales, all 1-4. The SAT has three 1-4 scales that are not summed, and UK tests such as A and AS Levels have multiple traits, usually four to six, that are always scored on scales that are multiples of 1-5 levels.
- **Most of the assessments, and all of the assessments that focused on the primary and secondary years/grades, allowed students access to dictionaries and, in some cases, grammar checkers or thesauri.** Some of the assessments are now on networked computers or tablets that include standard word processing applications with spell-checkers or dictionaries and other tools for writing.

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 large-scale 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 2011-2016 are displayed in Appendix I (see online report).

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 (see online report).

- 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.

Marking Criteria and supplementary materials. Moreover, for analysis, not all the criteria will be discussed nor in the exact order they appear in the guide.

TABLE 3: CORRELATION MATRIX — NAPLAN MARKING CRITERIA

	Audience	Text structure	Ideas	Persuasive devices	Vocabulary	Cohesion	Paragraphing	Sentence structure	Punctuation	Spelling
Audience	1									
Text structure	0.849	1								
Ideas	0.854	0.785	1							
Persuasive devices	0.852	0.807	0.799	1						
Vocabulary	0.842	0.770	0.795	0.800	1					
Cohesion	0.794	0.744	0.747	0.746	0.767	1				
Paragraphing	0.809	0.812	0.756	0.793	0.747	0.733	1			
Sentence structure.	0.840	0.796	0.793	0.776	0.785	0.775	0.768	1		
Punctuation	0.703	0.680	0.650	0.636	0.629	0.657	0.633	0.737	1	
Spelling	0.831	0.778	0.794	0.782	0.761	0.706	0.757	0.788	0.682	1

Source: Australian Curriculum Assessment and Reporting Authority (2017). *An Evaluation of Automated Scoring of NAPLAN Persuasive Writing: Technical Report*. Sydney p24 [Measurement Inc Report p21]

www.nap.edu.au/docs/default-source/default-document-library/technical-report_an_evaluation_of_automated_scoring_of_naplan_persuasive_writing.pdf?sfvrsn=2.

TABLE 4: PERCENT SHARED VARIANCE (CORRELATION COEFFICENT) — NAPLAN MARKING CRITERIA

	Audience	Text structure	Ideas	Persuasive devices	Vocabulary	Cohesion	Paragraphing	Sentence structure	Punctuation	Spelling
Audience	100.0%									
Text structure	72.1%	100.0%								
Ideas	72.8%	61.6%	100.0%							
Persuasive devices	72.5%	65.1%	63.8%	100.0%						
Vocabulary	70.8%	59.3%	63.1%	64.0%	100.0%					
Cohesion	63.0%	55.4%	55.8%	55.7%	58.8%	100.0%				
Paragraphing	65.5%	65.9%	57.1%	62.9%	55.8%	53.7%	100.0%			
Sentence structure	70.5%	63.3%	62.9%	60.1%	61.6%	60.1%	58.9%	100.0%		
Punctuation	49.5%	46.2%	42.3%	40.4%	39.5%	43.2%	40.1%	54.3%	100.0%	
Spelling	69.0%	60.5%	63.1%	61.1%	57.9%	49.8%	57.3%	62.1%	46.6%	100.0%

Source: Australian Curriculum Assessment and Reporting Authority (2017). *An Evaluation of Automated Scoring of NAPLAN Persuasive Writing: Technical Report*. Sydney p24 [Measurement Inc Report p21]

www.nap.edu.au/docs/default-source/default-document-library/technical-report_an_evaluation_of_automated_scoring_of_naplan_persuasive_writing.pdf?sfvrsn=2.

The major problem with the Audience criterion is that the NAPLAN prompt does not define any audience, making it impossible for the student writer to mould the script for a real or hypothesised audience. Instead, the marking criterion defines audience not as a specific type of person or persons, but as a generic reader, and the Audience Criterion as "The writer's capacity to orient, engage, and persuade the reader" (p6). Yet, orienting the reader approximates another criterion, "cohesion", which is defined as "The control of multiple threads and relationships across the text, achieved through the use of referring words, ellipsis, text connectives, substitutions, and word associations." In addition, the phrase "persuade the reader", implies the category Persuasive devices: "The use of a range of persuasive devices to enhance the writer's position and persuade the reader." Moreover, persuasion is also included in description of Ideas: "The selection, relevance and elaboration of ideas for a persuasive argument." Finally, although there is no reference to vocabulary, in practice, it also correlates very strongly with these other four criteria in the sample marking scripts. The very close relationship among these four variables is displayed in the correlation matrix in Table 5.

TABLE 5: CORRELATIONS OF FIVE INTERRELATED MARKING CRITERIA

	Audience	Ideas	Persuasive devices	Vocabulary	Cohesion
Audience	X				
Ideas	0.9688	X			
Persuasive devices	0.9427	0.9226	X		
Vocabulary	0.9305	0.9367	0.8974	X	
Cohesion	0.9668	0.9431	0.9154	0.9320	X

Source: Training script scores, NAPLAN 2017 Persuasive Writing Marking Guide

Albeit that these statistics derive from the small and artificial sample of the Marking Guide, training samples are what define the scoring of markers. Given that the markers are seeing correlations of 0.9 and above among these variables, there is a strong chance that rather than measuring five different attributes, these marks are measuring a single one.

Audience is also used inconsistently. In the marking criteria for audience, one of the bullet points is "takes readers' values and expectations into account". This is a clear reference to moulding a text for an audience, even though there is no audience specified. However, in the commentary for the script "They try to break out," (p37) the justification for a mark of 3 out of 6 is, "Argument is clear and supported with some evidence. The distinction between cages and zoos is made clear through reasons presented." Here, audience becomes synonymous with logical argument.

Ideas are defined as "the selection, relevance, and elaboration of ideas for a persuasive argument." The rubric appears to try to differentiate assertions from facts. The annotations on the training scripts, however, appear to be inconsistent and arbitrary. In some scripts, assertions such as "video games improve eyesight" (p74) are taken as solid substantiating evidence. In other scripts, however, statements such as "dogs love human attention" (p49) are considered mere assertions for the ideas criterion even though such statements too have some scientific standing (Coren, 2012). However, on page 49 of the Persuasive Marking Guide (as well as in the Marking Guide's Glossary, p86), the same phrase is described as an effective persuasive device.

Vocabulary is defined as "the range and precision of contextually appropriate language". Among the categories and examples are "modal adjectives and adverbs" such as definitely; "precise word groups" such as a "positive impact on society"; and nominalisations such as likelihood (no examples of nominalisations from verbs were given). Well-regarded style guides, however, caution against nominalisations from verbs (Pinker, 2014; Strunk and White, 1979; Williams and Bizup, 2017).

Paragraphing is not as much concerned with the overall organisation of the script, which is covered in Text Structure, as it is with the formation of paragraphs following the prescribed form of the five-paragraph essay, particularly with each paragraph beginning with a topic sentence. However, composition research has largely been critical of the five-paragraph essay (Brannon et al, 2008; Nunnally, 1991; Wesley, 2000). Moreover, it has long been known that in professional expository writing fewer than 20 per cent of paragraphs begin with a topic sentence (Braddock, 1974).

Spelling is by far the most curious criterion. As mentioned previously, similar tests now allow students access to dictionaries, and when writing online, access to spell-checkers and sometimes access to thesauri and grammar-checkers. Due to the ubiquity and effectiveness of spell-checkers, most spelling errors are caught. Even some homonyms such as there own are caught in Microsoft Word through the grammar checker. Yet in the NAPLAN essay, spelling accounts for six marks out of a total of 48 or 12.5 per cent, even though the Language Conventions portion that precedes the essay contains a substantial spelling section calibrated to students' specific Years (Australian Curriculum Assessment and Reporting Authority, 2016).

The calculation of the spelling mark 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 (at the end of Appendix K; see online report) categorising approximately one thousand 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 2. To attain a mark of 6, a student must spell all words correctly and include at least 10 difficult words and some challenging words or at least 15 difficult words.

This formulation is extremely difficult for markers, who may have only 27 seconds to check the Spelling reference list for all possible difficult and challenging words in a script. However, such a task is perfectly suited for a computer algorithm. Unfortunately for computers, this formulation of spelling, especially in the twenty-first century, has little, if anything, to do with measuring writing ability. And, indeed, the best way to prepare for this part of the test is for students to memorise hundreds of words and then insert them into the essay, regardless of whether they are the most appropriate words to convey the intended meaning. A knowledge of how to spell words correctly, or at least, how to quickly access the correct spelling of a word, is definitely a component of the writing construct. Memorising a spelling list is not.

Marking Guide glossary

The glossary at the end of the Marking Guide (pp86-97) and displayed in Appendix K (see online report) explains and expands the terminology used in the marking rubrics. Most of its information is elementary, trivial, and, in several places, incorrect. "Section I: Persuasive Devices" begins with a discussion of classical rhetoric. It asserts that ethos is "persuading by appealing to readers' values". That statement is incorrect. Beginning in Chapter 2 of Book I of *On Rhetoric,* Aristotle lists the three types of proof, *pisteis*, that are not pre-existing but created by the art, *techne*, of the rhetor, *ethos, pathos,* and *logos* (Aristotle, 2007; Aristotle, 1926; Garver, 1994). *Ethos* in classical Greek means character, and Aristotle's notion of an argument from ethos meant the character of the speaker created in the speech: such as *phronësis*, practical wisdom, *aretë*, virtue, and *eunoia*, good will (Book II, chapter 1). For Aristotle, ethos is created within the speech. For other Greek rhetoricians such as Isocrates, ethos also consists of the person of the speaker and his past achievements (Jarratt, 1998; Kennedy, 1994).

The mistaken notion that rhetorical ethos has to do with the values of the audience derives from the modern terms "ethics" and the modern use of "ethos" as the shared values of a community or culture. Both terms derive etymologically from the ancient Greek *ethos*, but their meanings have diverged substantially from the original.

Curiously, the explanation of *logos* as "appeal to reason" such as using "qualified measured statements" comes close to Aristotle's notion of ethos, but has little to do with proof through *logos*. The modern term logic, of course, derives from *logos* and is very close to Aristotle's notion of it. For Aristotle, proof through *logos* is proof through the syllogism or its abbreviated form, the enthymeme. The glossary's treatment of classical rhetoric is most absurd is its explanation of "Pathos — appeal to emotion". Aristotle devotes a substantial part of the second book of *The Art of Rhetoric* to persuasion through emotion and never does he include anything like "appeal to spurious authority".

The list of persuasive devices is sometimes equally arbitrary. As mentioned previously, an "authoritative statement" is given as an example. The use of the conditional mood is also listed as a persuasive device. Further strategies include underlining, bolding, and exclamation marks along with sarcasm, hyperbole, and expressions of personal opinion indicated by phrases such as "I think" and "In my opinion".

The rest of the glossary is fairly pedestrian except for the category "Referring Words", which "maintain continuity and avoid repetition". Included in this list are articles and quantifying determiners. Both articles and quantifying determiners, but especially articles, have nothing to do with continuity and repetition, which are semantic categories. Articles have no meaning outside of a theory of syntax.

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.

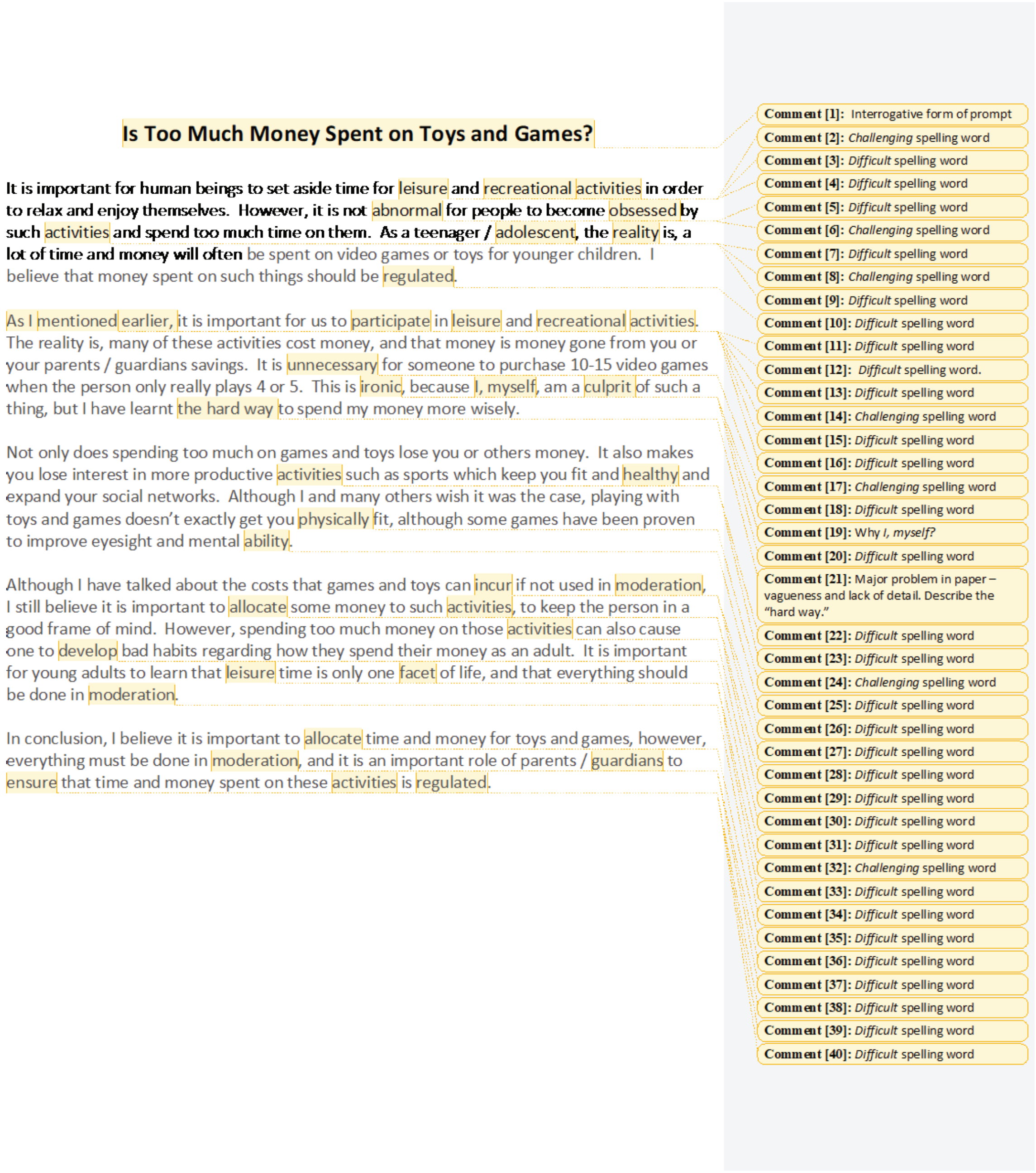
The most effective way to support the above assertion is to examine the exemplar script in the *Persuasive Writing Marking Guide* from at least 2012 to the present. By "exemplar script", I mean a perfect or near perfect script included in the training sample. In the case of the *NAPLAN Persuasive Writing 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 (see online report). 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.

FIGURE 2: EXEMPLAR PERSUASIVE SCRIPT WITH DIFFICULT AND 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 (see online report).

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.

FIGURE 3: 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.

Process

Before I even start outlining some of the basic questions, it is important to discuss who is in the room, that is, who should be involved in formulating the questions and developing possible answers. Of course, government ministers and their associates need to be there. However, I would argue that equally important for developing writing assessments is that the room be heavily populated with teachers of writing and professional writers, such as journalists. It might be profitable to include some parents and perhaps a few recent students. Maybe one or two business persons whose work involves writing should also attend. The one group whose presence should be severely limited is, psychometricians, the educational testers. One or two can be in the room but they should largely remain silent.

The role of psychometricians should be limited to technical issues, and teachers and writers should constitute the final authority on issues of validity and reliability. To paraphrase Carl Campbell Brigham, the American psychologist who created the Scholastic Aptitude Test (SAT)[1], it is probably simpler to teach cultured persons testing than to give testers culture (Brigham, 1937)[2]. Similarly, those who teach and professionally engage in writing, given the basic parameters needed for reliable assessments, can determine the specific goals of an assessment, design it to meet those goals, create the metrics for evaluation, and conduct the actual evaluation with much greater validity than writing assessments created by psychometricians.

The next step is to ask the simple question, "Why are we assessing?" Earlier, I reported the current use of NAPLAN:

1. as an indicator of school performance for use by school administrators in improving schools and for parents in selecting schools;
2. as a snapshot of national student achievement that can be used for longitudinal comparison; and
3. as formative feedback to students on their skills in literacy.

As in any reformulation process, we need to ask whether these are the right purposes. And this question should involve much research and discussion. One possible suggestion is to ask whether

[1] Brigham is a controversial figure. His early research involving American World War I intelligence tests provided the basis for the United States Immigration Act of 1924 that severely limited immigration mainly to Northern Europeans. In 1930, however, better data caused him to retract the entirety of his earlier research, including his racist and anti-Semitic assertions. As Secretary of the College Entrance Examination Board, Brigham created the Scholastic Aptitude Test (SAT) in 1926, but in the 1930s he also repudiated it. He opposed the creation of the Educational Testing Service and the paraphrase is taken from an article proposing an alternative testing organisation.

[2] The actual quotation is "it is probably simpler to teach cultured men testing than to give testers culture". It has been modified to remove the sexist language, which is an artefact of the period.

achievement as measured by an instrument such as NAPLAN is the best measurement for parents to use in evaluating schools. Recent research makes a convincing case that "effect", the improvement in abilities students achieve while at a school, is a much more relevant measure for selecting schools (Abdulkadiroglu, Pathak, Schellenberg and Walters, 2017). Individual school reports, for example, could be based on mean longitudinal improvement; how much, on average, students' scores improved during their stay at the school. Part of the general discussion, then, should be on the best ways to obtain and report data for specific and clearly defined purposes, including serious consideration of universal but locally based assessment.

Development should begin with experimentation, followed by independent evaluation of each experiment. Every step of development should be transparent by being matters of public record. Good engineering practice means there may be a need to reformulate specific parts of the practice several times. The central design principle should be constantly working toward the alignment of curriculum, teaching practice and assessment. Finally, tests should be constantly evolving to better serve teaching and learning. Set tests are helpful measures for longitudinal development, but they should not remain static at the expense of classroom effectiveness.

One Vision

I present my own possible scenario. It is not even a proposal. Think of it as a mental exercise to begin a discussion. If we want to test writing, we should test all aspects of it, including reading, prewriting, writing, revision and 21st century editing. That means including one or more set readings given to students ahead of the test, as is the case in the UK. There could even be several set readings, giving students a choice, as is the case in British Columbia's reading tests.

It has long been accepted knowledge in writing assessment that a single essay is an unreliable measure of writing ability. At least two essays, preferably in two different genres, are necessary. (Breland, Bridgeman and Fowles, 1999; Diederich, 1974; White, 1994). Furthermore, Diederich (1974) argues that the essays need to be separated in time, the minimum separation being morning and afternoon. Consequently, I propose a day-long test with two-and-a-half hour tests, one in the morning and one in the afternoon. Like the ACT Scaling Test, each script should be limited to 600 words, thereby emphasising revision and discounting length. Because the tests will take so much time, I suggest that instead of being administered four times, they be officially administered as summative assessments only twice (in Year 5 and in Year 9), following Peter Elbow's suggestion to assess less but assess better (1996). There would be, however, ample opportunities to employ versions of these instruments in classrooms as formative assessments.

Like the National Assessment of Educational Progress (NAEP) tests, there would be a rotation among the three genres with narrative favouring Year 5 and persuasive favouring Year 9. Because the emphasis should be on formative assessment, I favour the 6 + 1 Analytic Writing Continuum developed and refined by the US' National Writing Project. Markers first give a holistic mark and then mark in six categories, all on a 1-6 scale:

- Content, including the quality and clarity of ideas and meaning
- Structure
- Stance
- Sentence Fluency
- Diction (language, vocabulary, word choice)
- Conventions.

What is remarkable is the similarity of these categories to those of other multi-trait scoring systems. "stance" is close to the more familiar category of "voice", and Diederich's original formulation of the factor “flavour” (Diederich et al, 1961). Interestingly, this category although almost universal in most multi-trait marking schemes, is absent in NAPLAN marking, even though it comes closest to approximating the actual Aristotelian notion of *ethos.*

I have found the Analytic Writing Continuum to be remarkably robust. I have seen it used effectively marking Grade 3 essays and have employed it myself in a study of undergraduate writing at a major American research university. For use in NAPLAN, the exact definition of traits and the selection of marking scripts for training and the training guide should be led by teachers, with, at first, maybe some guidance from the National Writing Project. As is the case in British Columbia, NAPLAN marking could be local but moderated, while technology could be used to merge local and national marking. The marking sessions would take much more time. However, I have discovered that marking with the Analytic Writing Continuum and engaging in serious conversations about papers provides extremely effective and worthwhile in-service training. One of the most attractive features of the Analytic Writing Continuum is how effectively it has been used in the classroom as a tool for peer review and as an aid to revision (M.A. Smith and Swain, 2017). By learning and internalising the Analytic Writing Continuum, students become better writers, and the teaching becomes the testing. The principal goal of such an approach is to promote alignment of the curriculum, classroom pedagogy, and all forms of assessment, that is, to test to the teaching. If students consider classroom exercises and outside assessments to be indistinguishable, and both reflect the curriculum, then assessments reinforce teaching and learning rather than possibly subverting them.

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.

LES PERELMAN BIOGRAPHY

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 Communication-Intensive 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 2012-13, 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.*

WORKS CITED

Abdulkadiroglu, A, Pathak, P, Schellenberg, J, and Walters, C (2017) *Do Parents Value School Effectiveness?* Cambridge, MA. Retrieved from *www.nber.org/papers/w23912.pdf.*

Adler-Kassner, L, and Wardle, E.A. (Eds) (2016) *Naming what we know : threshold concepts of writing studies.* Boulder: Uinversity of Colorado Press.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing (US). (2014) *Standards for educational and psychological testing.* Washington DC: Joint Committee on Standards for Educational and Psychological Testing (US).

American Psychological Association. (1966) Standards for educational and psychological tests and manuals. Washington DC: American Psychological Association.

Ananda, S, and Rabinowitz, S (2000) *The high stakes of high-stakes testing.* San Francisco: WestEd. Retrieved from *files.eric.ed.gov/fulltext/ED455254.pdf.*

AQA. (2014). *A-level English language* (7702/2) *Paper 2: Language Diversity and Change.* Retrieved from *filestore.aqa.org.uk/resources/english/AQA-77022-SMS.PDF.*

Aristotle. (1926) *The art of rhetoric.* (J.H. Freese, Ed) Cambridge MA: Harvard University Press.

Aristotle. (2007) *On rhetoric : a theory of civic discourse.* (G.A. Kennedy, Ed) (2nd ed) New York: Oxford University Press.

Australian Curriculum Assessment and Reporting Authority. (nd) *English sequence of achievement.* Retrieved February 1, 2018, from *docs.acara.edu.au/resources/English_Sequence_of_achievement.pdf.*

Australian Curriculum Assessment and Reporting Authority. (2010) *2010 NAPLAN Narrative Writing Marking Guide*. Sydney: ACARA. Retrieved from *www.nap.edu.au/_resources/2010_Marking_Guide.pdf.*

Australian Curriculum Assessment and Reporting Authority. (2012) *2012 NAPLAN Persuasive Writing Marking Guide.* Sydney: ACARA. Retrieved from *www.nap.edu.au/_resources/2012_Marking_Guide.pdf.*

Australian Curriculum Assessment and Reporting Authority. (2013) *2013 NAPLAN Persuasive Writing Marking Guide.* Sydney. Retrieved from *www.nap.edu.au/_resources/Amended_2013_Persuasive_Writing_Marking_Guide_-With_cover.pdf.*

Australian Curriculum Assessment and Reporting Authority. (2016) *The tests*. Retrieved February 8, 2018, from *www.nap.edu.au/naplan/the-tests.*

Australian Curriculum Assessment and Reporting Authority. (2017a) *2017 NAPLAN Persuasive Writing Marking Guide.* Sydney: ACARA. Retrieved from *www.vcaa.vic.edu.au/Documents/naplan/schools/2017/Writing_Marking_GuideforDS.pdf.*

Australian Curriculum Assessment and Reporting Authority. (2017b) *Achievement in reading, writing, language conventions and numeracy national report for 2017.* Sydney. Retrieved from *www.nap.edu.au/docs/default-source/default-document-library/naplan-national-report-2017_final_04dec2017.pdf.*

Balf, T. (2014, March 6) *The story behind the SAT overhaul.* New York Times Magazine. Retrieved from *www.nytimes.com/2014/03/09/magazine/the-story-behind-the-sat-overhaul.html.*

Bellamy, P.C. (2001) *Research on writing with the 6+1 traits.* Portland: Northwest Regional Educational Laboratory. Retrieved from *citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.3088&rep=rep1&type=pdf.*

Berwick, R.C, Friederici, A.D, Chomsky, N, and Bolhuis, J.J. (2013) *Evolution, brain, and the nature of language.* Trends in Cognitive Sciences, 17(2), 89–98.

Braddock, R. (1974) *The frequency and placement of topic sentences in expository prose.* Research in the Teaching of English, 8(3) 287–302.

Brannon, L, Courtney, J. P, Urbanski, C. P, Woodward, S. V, Reynolds, J. M, Iannone, A. E, Kendrick, M. (2008) *The five-paragraph essay and the deficit model of education.* The English Journal, 98, 16–21.

Breland, H.M, Bridgeman, B, and Fowles, M.E. (1999) *Writing assessment in admission to higher education: review and framework.* College Board Report No 99-3. GRE Board Research Report No 96-12R. New York: College Entrance Examination Board. Retrieved from *eric.ed.gov/?q=College+Board+99-3&id=ED563111.*

Brigham, C.C. (1937) *The place of research in a testing organisation.* School and Society, 46 (1198) 756–759.

British Columbia Ministry of Education. (2013) *Foundation Skills Assessment 2013 monitoring report.* Retrieved from *www2.gov.bc.ca/assets/gov/education/kindergarten-to-grade-12/teach/pdfs/assessment/training/2013_fsa_monitoring_report.pdf.*

British Columbia Ministry of Education. (2016) *Foundation Skills Assessment administration instructions 2017.* Victoria: British Columbia Ministry of Education. Retrieved from *www2.gov.bc.ca/assets/gov/education/administration/kindergarten-to-grade-12/assessment/fsa-administration-instructions-eng.pdf.*

British Columbia Ministry of Education. (2017a) *2017 Foundation Skills Assessment: Description and specifications.* Retrieved from *www2.gov.bc.ca/assets/gov/education/administration/kindergarten-to-grade-12/assessment/fsa_description_specification_june2017.pdf.*

British Columbia Ministry of Education. (2017b) *Collaboration activity Grade 4 — Sample.* Retrieved from *www2.gov.bc.ca/assets/gov/education/administration/kindergarten-to-grade-12/assessment/grade-4-collaboration-activities-sample.pdf.*

British Columbia Ministry of Education. (2017c) *Foundation Skills Assessment scoring guide: Grades 4 and 7.* Victoris: British Columbia Ministry of Education. Retrieved from *www2.gov.bc.ca/assets/gov/education/administration/kindergarten-to-grade-12/assessment/fsa-scoring-guide-en.pdf.*

British Columbia Ministry of Education. (2017d) *New Foundation Skills Assessment.* Retrieved January 23, 2018, from *curriculum.gov.bc.ca/assessment-reporting/new-foundation-skills-assessment.*

Cheng, L, Fox, J, and Zheng, Y. (2007) *Student accounts of the Ontario secondary school literacy test: A Case for Validation.* Canadian Modern Language Review, 64(1), 69–98. *doi.org/10.3138/cmlr.64.1.069.*

Cochran-Smith, M. (1991) *Word processing and writing in elementary classrooms: a critical review of related literature.* Review of Educational Research, 61(1), 107–155.

College Board. (2015) *Test specifications for the redesigned SAT.* New York: College Board. Retrieved from *collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf.*

College Board. (2017) *SAT Essay scoring.* Retrieved September 15, 2017, from *collegereadiness.collegeboard.org/sat/scores/understanding-scores/essay.*

College Board. (2018) *SAT Essay.* Retrieved January 24, 2018, from *collegereadiness.collegeboard.org/sat/inside-the-test/essay.*

Connors, R.J and Lunsford, A.A. (1988) *Frequency of formal errors in current college writing, or Ma and Pa Kettle do research.* College Composition and Communication, 39(4), 395. *doi.org/10.2307/357695*

Coren, S. (2012, January) *Do dogs love people more than they love other dogs?* Psychology Today. Retrieved from *www.psychologytoday.com/blog/canine-corner/201201/do-dogs-love-people-more-they-love-other-dogs.*

Deane, P. (2013) *On the relation between automated essay scoring and modern views of the writing construct.* Assessing Writing, 18(1), 7–24. *doi.org/10.1016/j.asw.2012.10.002.*

Diederich, P.B. (1974) *Measuring growth in English.* Urbana IL: National Council of Teachers of English.
Diederich, P. B, French, J. W and Carlton, S. T. (1961) *Factors in judgements of writing ability.* ETS Research Bulletin Series (Vol. 1961). Princeton NJ. Retrieved from *doi.wiley.com/10.1002/j.2333-8504.1961.tb00286.x.*

Elbow, P. (1996) *Writing assessment: Do it better, Do it less.* In W. Lutz, E.M. White and S. Kamusikiri (Eds), The politics and practices of assessment in writing. New York: Modern Language Association.

Elliot, N. (2005) *On a scale : a social history of writing assessment in America.* New York: Peter Lang.

FairTest. (2007) *The new SAT: A better test or just a marketing ploy?* FairTest. Retrieved January 6, 2018, from www.*fairtest.org/new-sat-better-test-or-just-marketing-ploy.*

Fischer, S. R. (2001) *History of writing.* Islington, UK: Reaktion Books.

Flower, L and Hayes, J.R. (1981) *A cognitive process theory of writing.* College Composition and Communication, 32(4), 365.

Garver, E. (1994) *Aristotle's Rhetoric: An art of character.* University of Chicago Press.

Godshalk, F.I, Swineford, F, Coffman, W. (1966) *The measurement of writing ability.* Princeton: College Entrance Examination Board. Retrieved from *eric.ed.gov/?id=ED029028.*

Gowers, E and Fraser, B. (1973) *The complete plain words (Revised).* London: HM Stationery Office.

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

Ireland, J. (2017, September 17) *National tests for year one students under government plan.* Sydney Morning Herald. Retrieved from *www.smh.com.au/federal-politics/political-news/national-tests-for-year-one-students-under-government-plan-20170917-gyj108.html.*

Jarratt, S.C.F. (1998) *Rereading the sophists : classical rhetoric refigured.* Southern Illinois University Press.

Kane, M.T. (2013) *Validating the interpretations and uses of test scores.* Journal of Educational Measurement, 50(1), 1–73.

Kennedy, G.A. (1994) *A new history of classical rhetoric.* Princeton NJ: Princeton University Press.

Lam, F.S and Pennington, M.C. (1995) *The computer vs the pen: A comparative study of word processing in a Hong Kong secondary classroom.* Computer Assisted Language Learning, 8(1), 75–92.

Liu, M, Moore, Z, Graham, L, and Lee, S. (2002) *A look at the research on computer-based technology use in second language learning.* Journal of Research on Technology in Education, 34(3), 250–273.

Lloyd-Jones, R. (1977) *Primary trait scoring. Evaluating writing: describing, measuring, judging.* Retrieved from *files.eric.ed.gov/fulltext/ED143020.pdf#page=43.*

Luce-Kapler, R and Klinger, D. (2005) *Uneasy writing: The defining moments of high-stakes literacy testing.* Assessing Writing, 10(3), 157–173.

Lunsford, A.A and Lunsford, K.J. (2008) *Mistakes are a fact of life: A national comparative study.* College Composition and Communication, 59(4), 781–806.

Malady, M.J.X. (2013, October) *We are teaching high school students to write terribly: The many problems of the SAT's essay section. Slate.* Retrieved from *www.slate.com/articles/life/education/2013/10/sat_essay_section_problems_with_grading_instruction_and_prompts.html.*

McCurry, D. (2010) *Can machine scoring deal with broad and open writing tests as well as human readers?* Assessing Writing, 15(2), 118–129.

Mislevy, R.J. (2018) *Sociocognitive Foundations of Educational Measurement.* London: Routledge.

Murray, D.M. (1982) *Learning by teaching: selected articles on writing and teaching.* Portsmouth NH: Boynton/Cook.

National Assessment Governing Board. (2010) *Writing framework for the 2011 national assessment of educational progress.* Washington. Retrieved from *www.nagb.gov/content/nagb/assets/documents/publications/frameworks/writing/2011-writing-framework.pdf.*

National Center for Educational Statistics. (2017) *Timeline for national assessment of educational progress (NAEP) assessments from 1969 to 2024.* Retrieved January 14, 2018, from *nces.ed.gov/nationsreportcard/about/assessmentsched.asp.*

National Center for Education Statistics. (2011) *The nation's report card: Writing 2011.* Washington DC. Retrieved from *nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf.*

National Center for Education Statistics. (2016) *NAEP — What does the NAEP writing assessment measure?* Retrieved January 6, 2018, from *nces.ed.gov/nationsreportcard/writing/whatmeasure.aspx.*

Newkirk, T. (1981) *Barriers to revision.* Journal of Basic Writing, 3(3), 50–61. Retrieved from *wac.colostate.edu/jbw/v3n3.*

Nunnally, T.E. (1991) *Breaking the five-paragraph-theme barrier.* The English Journal, 80(1), 67–71. *doi.org/10.2307/818100.*

Orwell, G. (1954) *A collection of essays.* New York: Harcourt Brace Jovanovich.

Perelman, L. (nd) *Dr Perelman's SAT essay writing tips.* Retrieved February 11, 2018, from *www.actoutagainstsat.com/essay-tips.pdf.*

Perelman, L. (1999) *The two rhetorics: Design and interpretation in engineering and humanistic discourse.* Language and Learning Across the Disciplines, 3(2), 64–82. Retrieved from *wac.colostate.edu/llad/v3n2/perelman.pdf.*

Perelman, L. (2012) *Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES).* In Bazerman, C; Dean, C; Early, J; Lunsford, K; Null, S; Rogers, P; Stansell (Eds) *International advances in writing research* (pp121–131) Fort Collins, Colorado: The WAC Clearinghouse and Parlor Press. Retrieved from *wac.colostate.edu/books/wrab2011/chapter6.pdf.*

Perelman, L. (2014) *When "the state of the art" is counting words.* Assessing Writing, 21.

Persky, H. (2012) *Writing Assessment in The Context of The National Assessment of Educational Progress.* In N. Elliot; L. Perelman (Eds) *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp69–86). New York: Hampton Press.

Pinker, S. (2014) *The sense of style : the thinking person's guide to writing in the 21st century!* New York: Viking.

Pratt, M.L. (1977) *Toward a speech act theory of literary discourse.* Bloomington: Indiana UP.

Purves, A.C. (1992) *Reflections on research and assessment in written composition.* Research in the Teaching of English, 26, 108–122.

Robinson-Staveley, K and Cooper, J. (1990) *The use of computers for writing: effects on an English composition class.* Journal of Educational Computing Research, 6(1), 41–48. *doi.org/10.2190/N3WK-KC2Q-DVGD-7F0B.*

Russell, M. (1999) *Testing On Computers* Education Policy Analysis Archives, 7 (20).

Singer, N.R and LeMahieu, P. (2011) *The Effect of scoring order on the independence of holistic and analytic scores.* Journal of Writing Assessment 4(1) Retrieved from *journalofwritingassessment.org/article.php?article=51.*

Slomp, D.H. (2008) *Harming not helping: The impact of a Canadian standardized writing assessment on curriculum and pedagogy.* Assessing Writing, 13(3), 180–200. *doi.org/10.1016/J.ASW.2008.10.004.*

Smarter Balanced Assessment Consortium. (nd) *Smarter Balanced sample items.* Retrieved January 21, 2018, from *sampleitems.smarterbalanced.org.*

Smarter Balanced Assessment Consortium. (2015) *Content specifications for the summative assessment of the common core state standards for English language arts and literacy in history/social studies, science, and technical subjects.* Los Angeles: Smarter Balanced Assessment Consortium. Retrieved from p*ortal.smarterbalanced.org/library/en/english-language-artsliteracy-content-specifications.pdf.*

Smarter Balanced Assessment Consortium. (2017) *Members and governance — Smarter Balanced Assessment Consortium.* Retrieved January 20, 2018, from *www.smarterbalanced.org/about/members.*

Smith, A. (2014, August 18) *Students score poorly in "confusing" NAPLAN writing test. Sydney Morning Herald.* Retrieved from *www.smh.com.au/nsw/students-score-poorly-in-confusing-naplan-writing-test-20140815-104liu.html.*

Smith, M.A and Swain, S.S. (2017) *Assessing writing, teaching writers : putting the analytic writing continuum to work in your classroom.* New York: Teachers College Press.

Sommers, N. (1980) *Revision strategies of student writers and experienced adult writers.* College Composition and Communication, 31(4), 378.

Sommers, N. (1981) *Intentions and revision.* Journal of Basic Writing, 3(3), 41–49. Retrieved from *wac.colostate.edu/jbw/v3n3.*

Sommers, N. (1982) *Revision.* Journal of Teaching Writing, 1(2). Retrieved from *journals.iupui.edu/index.php/teachingwriting/issue/view/78.*

Strauss, V. (2013, January 16) *Pearson criticized for finding test essay scorers on Craigslist.* The Washington Post. Retrieved from *www.washingtonpost.com/blogs/answer-sheet/wp/2013/01/16/pearson-criticized-for-finding-test-essay-scorers-on-craigslist.*

Strunk, W. and White, E.B. (1979) *The elements of style (4th ed).* New York: Pearson.

Swain, S.S and LeMahieu, P. (2012) *Assessment in a culture of inquiry: The story of the national writing project's analytical writing continuum.* In N. Elliot; L. Perelman (Eds) Writing Assessment in the 21st Century: Essays in Honor of Edward M. White (pp45–68).

Weiss, John. (1987) *The Golden Rule Bias Reduction Principle: A practical reform.* Educational Measurement: Issues and Practice, 6(2), 23–25. 3992.1987.tb00408.x.

Weiss, Joanna. (2014, March 14) *The man who killed the SAT essay.* Boston Globe. Retrieved from *www.bostonglobe.com/opinion/2014/03/13/the-man-who-killed-sat-essay/L9v3dbPXewKq8oAvOUqONM/story.html.*

Wesley, K. (2000) *The ill effects of the five paragraph theme.* The English Journal, 90(1), 57–60.

White, E.M. (1984) Holisticism. College Composition and Communication, 35(4), 400.

White, E.M. (1994) *Teaching and assessing writing: recent advances in understanding, evaluating, and improving student performance (2nd ed).* San Francisco: Jossey-Bass.

Williams, J. M. and Bizup, J. (2017) *Style : lessons in clarity and grace.* New York: Pearson.

Winerip, M. (2005, May 4) *SAT Essay Test rewards length and ignores errors.* New York Times. Retrieved from *www.nytimes.com/2005/05/04/education/sat-essay-test-rewards-length-and-ignores-errors.html.*

Notes

Notes

100 YEARS OF TEACHER UNITY
N.S.W.T.F.
QUI DOCET DISCIT
1918 - 2018

Published by
NSW Teachers Federation
23-33 Mary Street Surry Hills NSW 2010
AUSTRALIA
www.nswtf.org.au