A Robust and Generalisable Rubric Design Framework for Critical Thinking Assessment

Thesis submitted for the degree of Doctor of Philosophy

at the University of Leicester

by

Harry A. Layman

2021

Abstract

A Robust and Generalisable Rubric Design Framework for Critical Thinking Assessment

Harry A. Layman

This research has two primary goals. The first is to develop a useful framework for designing rubrics to improve the utility of feedback and the reliability of scoring for critical thinking assessments that use constructed response items. The second is to demonstrate and explore the practicality, effectiveness, strengths, and weaknesses of this approach as applied to some specific data sets.

The use of constructed response (CR) for educational assessment has been advocated for decades (Gulikers, Bastiaens and Kirschner, 2004; Palm, 2008; Wiggins, 1990). The primary benefits claimed are more authentic measurement and better feedback to students and teachers. More authentic measurement includes the notion that the construction of a response is more cognitively challenging and more direct evidence than the indirect evidence from selecting among predefined choices. Better feedback is generally limited in practice, however, when use of CR items relies on holistic scoring, generic rubrics, and a regimen of scorer training and calibration to attain consistent and generally valid measurement. The resulting broad, multifactor classification levels are unable to convey response-specific feedback (Bejar, 2017).

This research postulates the use of a rubric design framework for the creation of item-specific, content-centric rubrics for assessment items that have right and wrong, better and worse possible responses. The framework establishes a uniform mechanism for identifying essential elements of item responses, with explicit weights for varying degrees of correctness and completeness and standardised approaches to calculating overall scores, subscores, and scaled scores. Resultant score reports can provide explicit feedback to response elements present, absent, or less than complete that explicitly justify and explain score differences between responses. Successful use of this rubric design framework promises CR assessment for critical thinking or argumentative writing items that will have score reports able to provide detailed

feedback to students and defensible scoring outcomes, with the potential for improved interrater reliability.

Acknowledgements

First, I want to express my gratitude to my family and friends for enduring more tedium than they deserve, especially Patti, Andy, Harry, Colleen, and James.

I am indebted also to the scholars, educators, and technologists who have helped, instructed, and encouraged me on this journey, including Gerald Melican, Wayne Patience, Randy Bennett, Wim J. van der Linden, and Thomas Proctor. Special thanks go to Elijah Mayfield for tutoring me in the techniques of machine learning for text and to Yury Matveev for inspirational and pivotal Python tutoring and support.

Thanks also to Maura O'Riordan, Roscio Cisneros Beltran, Xaviera Gonzalez-Wagener and Jennifer Bezirium for invaluable assistance with reviewing rubrics and scoring essays.

I would like to thank and acknowledge Elijah Mayfield and Turnitin LLC (now Part of Advance) for signing off on my request, through UCI's Carol Booth Olson, Professor Emerita, and Huy Quoc Chung, PhD of the California Writing Project, to share the writing program responses and data.

And most of all, thanks to Glenn Fulcher, my adviser, for pointers to new knowledge, introduction to new authorities, and encouragement and guidance that always made sense—and an unfailing belief that I could accomplish my goal.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Acronym	**Reference**
AW	Argumentative writing
CAE	Council for Aid to Education (cae.org)
CLA	Collegiate Learning Assessment from CAE
CR	Constructed response
CT	Critical thinking
ETS	Educational Testing Service of Princeton, New Jersey
GRE	Tests offered by ETS for graduate school admissions
IRR	Interrater reliability
HT	Harriet Tubman
MCQ	Multiple-Choice Question (see SR)
QWK	Quadratic Weighted Kappa (a version of Cohen's kappa)
RDF	Rubric Design Framework
SR	Selected response
WH	Winter Hibiscus

Chapter 1 Introduction

1.1 Background of the Study

For many decades, multiple-choice questions have been the item type of choice for standardised testing, for reasons of cost and practicality. Even as that preference has held, selected response (SR) has been criticised for not reflecting the cognitive processes and capabilities that are being evaluated (Wiggins, 1990); for not capturing higher level cognition (Liu, Frankel and Roohr, 2014); for not representing real-world problems (Birenbaum, 1996); and for decontextualising knowledge rather than demonstrating abilities of synthesis, analysis, and argumentation (Gulikers, Bastiaens, & Kirschner, 2004). This criticism has made at least some level of constructed response (CR) obligatory in most high-stakes exams.

The gradual increase in the use of CR for standardised testing over time has been enabled by decades of research and experimentation, which addressed initial challenges to reliability (due to low interrater agreement) and validity (Bejar, 2017, p. 570). The standardisation of CR scoring, as exemplified by Myers (1980) and further described in Baldwin, Fowles and Livingston (2005), insured sufficient CR reliability by establishing a standardised approach to holistic scoring that included standardised techniques for scorer training, scorer calibration, and scorer monitoring. Over the same period, the idea of validity, led by the work of Sam Messick at ETS, evolved beyond the trinitarian view of content, criterion, and construct validities to a unitarian view (Messick, 1980, 1989) that included the consequences of a test's use. Validity has since expanded to include a robust view of social values and consequences, including fairness (Kane, 2010).

Although the broadening of the notion of test validity and the increased reliability of CR items made way for greater adoption of CR, the ever-lower cost of multiple-choice testing combined with shorter test time per item, amplified the efficiency difference between the two item types. Calls for more authentic assessment, direct assessment that promotes learning, assessment as learning, performance tasks, and other forms of testing that involve student construction of a response rather than selecting from

predefined choices have never abated.[1] Many educational assessments continue to use at least a small number of CR items to enhance face validity, even where the contribution to measurement is minimal. For example, Zahner (2013) analysed the contribution of a single CR item (a 'performance task' requiring an extended constructed response) on an exam otherwise composed of 25 SR items. In addition, the two primary US college admissions exams currently consist entirely of SR questions (154 for the SAT and 215 for the ACT) and a single, optional, CR essay item (College Board, no date).

One area where CR items remain well established is in the assessment of more complex skills and higher level cognition, including critical thinking (CT) and argumentative writing (AW) (Liu, Frankel and Roohr, 2014). As some 95% of surveyed institutions identified CT skills as among the most important learning outcomes for their students (Association of American Colleges and Universities, 2011), this study focuses on the use of CR for these important educational measurements. A common denominator of skills assessed by CT assessments is the ability to make claims and cite supporting evidence (Liu, Frankel and Roohr, 2014; Jackson, Draugalis, Slack, & Zachry, 2002; and Lomask & Baron, 2003), a challenge on which this study focuses.

Most high-volume CT exams rely on multiple-choice questions and use a small number of CR items. As Butterworth and Thwaites (2013, pp. 342–343) noted, exams such as the Biomedical Admissions Test, the Cambridge Thinking Skills Assessment, Singapore's H2 Knowledge and Enquiry assessment, and the Theory of Knowledge portion of the International Baccalaureate all include content designed to measure CT skills. Many of these exams are used widely in such countries as the Netherlands, Spain, Malaysia, Singapore, and Thailand. In terms of number of items, allocated student time, or proportion of summative result scores, such exams still rely heavily on multiple-choice question (MCQ) items while including at least one CR item to

[1]In the longer historical perspective, SR was in fact the 'new type of testing' given great impetus by the work of Fredrick Kelly (1916), from which the push back toward performance testing was signaled as early as Kaulfers (1944), which called for actual language production rather than overreliance on SR in language achievement assessment in situations where safety or military effectiveness might be compromised.

obtain an authentic sample of student-produced work on which to base some portion of the assessment results.

1.2 The Problem With Existing Rubrics

CR items used in educational assessment routinely use generic rubrics and holistic scoring. Such items can achieve acceptable levels of reliability but are unable to provide useful, detailed feedback (Cumming, Kantor and Powers, 2002, and Bejar, 2017); can have minimally acceptable levels of reliability (Williamson, Xi and Breyer, 2012, p. 7); have scores that can be difficult to justify (Harsch & Martin, 2013 and Carr, 2020); and require time and effort that, from the students' perspective, yields little educational value (Cumming, Kantor & Powers, 2002; Nordrum, Evans, and Gustafsson, 2013).

When standardised testing has moved toward greater score detail by changing from a single, holistic score with 3–6 score points on a single scale to multiple dimensions for complex-cognition CR items, the result has often been the use of two or three subscores on an even smaller scale, typically 3 or 4 score points. (Examples include the essay scoring evolutions seen in the College Board's SAT admissions exam programme, ETS's GRE graduate school admissions test programme, and Graduate Management Admissions Council's graduate management school admission testing programme.) Attempts to provide subscores often show them to have both low reliability and high collinearity.

Two examples capture the struggle for useful and detailed CR scoring for CT skills:

- A recent AACU research study found that the percentage agreement in scoring was fairly low when multiple raters scored the same student work using the VALUE rubrics (Finley, 2012, cited in Liu, Frankel and Rohr, 2014). For example, the percentage of perfect agreement of using four scoring categories across multiple raters was only 36% when the CT rubric was applied (Liu, Frankel and Roohr, 2014).
- In a CAE report on their CLA+, which has a performance task to measure analytic reasoning, Zahner (2013) reported the reliability of the 60-minute CR section as only 0.43.

1.3 Purpose and Nature of the Study

The purpose of the study was to develop a rubric design framework (RDF) for use in developing and evaluating rubrics for CR items designed to measure CT skills. The RDF is based on a theoretical model of CR scoring, tailored to the needs of CT assessment. The theoretical model of scoring builds on the existing literature and practice for CR scoring, informed by current research and practice in CT scoring, and leverages recent research and consideration of the role of rubrics in CR assessment.

The study comprises a literature review of CR scoring in general, CT assessment in particular, including related research on AW scoring and the various kinds of CR rubrics used today for educational assessment in general. After a review of a comprehensive framework for describing rubric elements, this study proceeds to review these elements and identify key components of a CT rubric to be used for CR items, a preliminary "rubric design framework" or RDF for evaluation of CT rubrics (represented in Appendix J).

The work then proceeds in two phases, a development phase and an evaluation phase. In each phase, three distinct scenarios are considered. In the development phase the preliminary framework is used to create an initial item-specific, content-centred rubric explicitly based on the goals and objectives for CT assessment and constructed piece by piece to correspond to the proposed RDF for evaluation. This preliminary RDF is used to design and evaluate rubrics for three separate scenarios, each including complete data sets composed of (a) an item passage or pair of passages, (b) a challenge or question to be addressed based on the passage(s) for the item, (c) a holistic rubric that was used to evaluate the responses and the scoring the resulted from its application, (d) a set of two scores for each response by each of two graders created by application of the holistic rubric. During the development phase, a new RDF aligned rubric is defined to score the responses, and it is applied to 40 of the item responses in the initial data set that represent the full range of scores obtained by the holistic scoring. The new RDF scoring is then used to evaluate the rubric and guide potential changes.

To complete the development phase of the work, the results of the application of the preliminary rubric evaluation framework are analysed, both in terms of the overall impact on scoring with the RDF rubric as compared to the scoring with the holistic rubric, and in terms of comparative IRR for the two scoring activities. The RDF scoring is next examined in detail and some adjustments or extensions to the RDF and the specific rubrics are considered. The rubrics for the three scenarios are adjusted accordingly, and the Testing phase of the work begins.

In the testing phase, additional item responses from the original data set are scored with the improved rubrics in each of the three scenarios. The testing phase includes a more comprehensive set of scoring work and evaluation activity with 120 items for one scenario (the first scenario, with shorter responses averaging 100 words) and 80 items for the second and third scenarios (longer responses averaging over 300 words). The testing phase concludes with a review of the scoring results and a comprehensive analysis of the results as compared to holistic scoring for the same item responses and as compared to the earlier RDF results for change or improvement. A summative analysis of the overall performance of the new RDF rubrics is then presented to compare the RDF results with the holistic scoring and to address the original research questions that motivate the study.

1.4 Research Questions

Research questions addressed in the study are:

1. Can a generalised and flexible RDF for scoring CT items (as compared to generic, holistic rubrics) be successfully used to define item-specific, content-centric rubrics that can guide essay graders to provide

- useful feedback to students and teachers;
- nuanced scoring that makes the exercise a learning experience;
- explicit, defensible rationales for scoring outcomes; and
- better interrater reliability?

2. Are there aspects of scoring with item-specific, content-centric rubrics that work well or that make scoring easier or more efficient?

1.5 Outline of the Thesis

This thesis is composed of seven chapters.

Chapter 1 introduces the research problem, providing a general background to challenges with CR scoring generally and CT scoring with holistic rubrics in particular, and then crystalizing the problem that is the focus of this study, followed by a description of the nature and purpose the study, and an enumeration of the specific research questions of interest.

Chapter 2 reviews the literature in three areas: research relevant to the whole of CR scoring, literature concerning CT and AW assessment, and general research on the use of rubrics in educational assessment. Dawson's (2017) rubric dimensions are introduced.

Chapter 3 explicates a theoretical model of scoring, reviews the goals and objectives of CT assessment, and reviews in detail the 14 elements of Dawson's taxonomy of rubric elements. The chapter then proceeds to build a ten-part rubric design framework (RDF) for CT that recognises the particular nature of the challenges of CT assessment and that incorporates Dawson's taxonomy of terms to provide a comprehensive approach to the consideration of CT rubrics for CR items in educational measurement.

Chapter 4 describes the research questions and the research context and describes the data requirements for the study and the item and item response data that was selected for the study. The basic methodology for the study is described in terms of three scenarios (item–rubric combinations) and two phases (development and testing) devised for this study that are used in subsequent chapters to analyse and compare the effect of the proposed rubrics on scoring process and results as compared to scoring with holistic rubrics. Chapter four also describes the important scoring protocol used for the scoring work undertaken as part of this study, and describes the research in terms of instruments, participants, and processes. The instruments include the items, prompts, passages, and holistic rubrics that form the baseline for this study of a new

framework for defining and evaluating rubrics. Data sources, respondents, and grader information are also described.

Chapter 5 reviews the RDF framework and defines a preliminary RDF-based rubric for each scenario in three discrete scenarios. For each scenario, the holistic rubric is described, and a proposed new rubric is created using the RDF as a guide. After preliminary scoring with the RDF rubric, scoring results are presented and analysed in comparison with the original holistic scoring results for the same item responses. The RDF results are also analysed to identify potential challenges and ambiguities in the preliminary rubrics, issues are identified and specific adjustments to the rubric are proposed.

For each of the three scenarios, Chapter 6 then applies the revised RDF rubrics to a larger number of item responses. The results of this scoring are again analysed in comparison to the earlier preliminary RDF scoring results and with the holistic scoring for these same items in the new, larger item response set. Differences in interrater reliability (IRR) and scoring outcomes are described. Implications for research questions including feedback, IRR, and scoring relative to holistic rubrics and the specific improvements made to the RDF in the development phase are discussed.

Chapter 7 reviews the findings across the three scenarios and concludes the study with a discussion of the practical and theoretical implications and limitations of the study. Suggestions for further research and a summary of the conclusions follow.

Chapter 2 Literature Review

This literature review focuses on three related topics that directly support this study: scoring for CR items generally; current issues and ideas for CT assessment; and research in the area of rubrics, rubric design, and related topics that are important to scoring and are related to rubric structure and use. These threads will all be tied together in support of the development of an RDF, which is the subject of Chapter 3.

2.1 Constructed Response

CR is a generic term for an assessment item type that requires an examinee to create or generate a response, rather than select from a set of predefined choices. Its most familiar forms include essay questions and short answers, for which the expected responses are a narrative passage or a word or phrase, respectively. The essential difference is that CR item responses are constructed or generated by the examinee, rather than selected from choices. SR and CR can come in many forms, but the canonical examples are the single-correct-answer multiple-choice and the essay question. As more fully developed in Bennett (1991, p. 4), CR items as a category can also include responses that are performed and so are 'closely associated with performance and "authentic" assessment'.

Myers (1980) focused on holistic scoring of essays and helped established the foundational procedures in CR scoring, such as the use of what Myers called anchors to define or illustrate different scoring categories. Myers also defined operational approaches to solving scoring decisions when competing factors are under consideration for a singular, holistic score. Myers addressed primary trait scoring, analytical scoring, and discourse scoring; in all cases he defined a group-oriented approach to the scoring task in a social context that encourages and supports consistency in rubric application. He documented collaborative scoring procedures that worked by building consensus among graders around anchor or prototype responses that exemplified quality levels for scores. He then defined processes whereby the group would use their consensus to help draw out the specific aspects of these exemplars that were useful in distinguishing different quality levels in the minds of his graders. This standardisation of operational procedures was credited with

enabling higher levels of IRR necessary for the use of CR items, albeit sparingly, in standardised assessment.[2]

Myers's (1980) work also established the importance of standardised procedures, instructions, exam context and environment, rules and test conditions, and anonymised response scoring. Exemplars and standardised procedures both remain relevant and important in CR scoring, including in this study, where I focus on the role of the rubric but do so within the context of CR scoring informed by these beneficial developments.

Procedures for scoring written items that focused on the assessment of writing provided a foundation for more generalised procedures for scoring performance assessments, which entailed written work products that were designed to provide authentic evidence of what an examinee knows and can do. The first sentence of the preface to Baldwin, Fowles and Livingston (2005, p. i) stated that the guidelines it contains were designed for a broad range of assessments such as CT and AW, as they fall squarely into the category of 'constructed-response questions, structured performance tasks, and other kinds of free-response assessments that ask the examinee to display certain skills and knowledge'. The guidelines for such scoring, like the earlier guidelines for scoring essays, attest to the rigour needed not just in defining what an assessment measures but also in the entire assessment enterprise, from defining the domain of knowledge and skills to be assessed to ensuring that an assessment is valid for its intended purpose. The concern for and emphasis on context, explicit enumeration of that is being measured, and concern for intended use all reflect the increasingly broad view of validity that holds true to this day.

The work of Baldwin, Fowles and Livingston (2005) captured the progress of the next decade and a half. They focused on assessment planning, writing assessment specifications and scoring specifications, defining the tasks and scoring criteria, pretesting, scoring, and assessment administration, providing a thorough and

[2] As noted in Bejar (2017, pp. 583–584), the path to present practices is more complex. Large-scale use of CR in the mid-1990s revealed reliability problems with CR scoring that led to a return of focus on MCQ assessment in K–12 standardized assessment for a few years, until improvements in the next decade (such as those noted by Baldwin, Fowles and Livingston, 2005) helped re-establish the modest level of CR use seen to this day in these assessments.

comprehensive guide that built on the work of prior decades. The paragraphs that follow further delineate additional concerns that have become essential elements of CR assessment for performance tasks that include CT and AW.

A major enhancement to standard assessment considerations highlighted by Baldwin, Fowles and Livingston's (2005) CR guidelines is the focus on externally facing processes designed to create transparency and trust with stakeholders, ranging from educators and administrators to policy experts, parents, and students. The guidelines also standardised a set of terms: *task*, where others might use words such as items, assignments, prompts, questions, problem, or topic; *response* for the performance or work to be evaluated, including an essay or an extended answer; *rubric* for scoring criteria, scoring guide, rating scale, and descriptors—or any framework used to evaluate responses; and *scorers* for raters, markers, readers, and so on. With the exception that this study will use the terms scorers and raters interchangeably, this study uses these terms as defined here throughout.

Baldwin, Fowles and Livingston's (2005) CR guidelines also contained a sharp focus on planning and design for every aspect of an assessment, emphasising the importance of purpose and intended use, which are at the heart of a comprehensive view of validity. These guidelines noted that the domain (content and skills) to be assessed as part of the test specification is to be called out with precision and that the demographics of the target population for the test—including academic background, grade level, professional goals, and so on—are important factors in the design of any assessment.

Similarly, assessment plans should address the need to collect validity evidence and ways to address issues of reliability (Baldwin, Fowles and Livingston, 2005). Validity needs to go beyond content coverage and include considerations of fairness and sufficient evidence of performance, including scoring that captures essential elements of performance. The test should also be delivered and experienced in a standardised way in every important aspect. Reliability appropriate for the test purpose also influences the number and range of tasks and the number of independent observations (independent scorers for each response).

Baldwin, Fowles and Livingston (2005, p. 3) noted also that

> a test taker's score should be consistent over repeated assessments using different sets of tasks drawn from the specified domain. It should be consistent over evaluations made by different qualified scorers. Increasing the number of tasks taken by each test taker will improve the reliability of the total score with respect to different tasks. Increasing the number of scorers who contribute to each test taker's score will improve the reliability of the total score with respect to different scorers.

The major improvements to assessment specifications and scoring specifications for CR assessment, including those most relevant to the focus on CT assessment as more fully described in the next section, are shown in Figure 2-1.

Figure 2-1 Highlights of Assessments and Scoring Specifications

In 'writing assessment specifications':

- The domain of knowledge and skills to be assessed should be precisely defined.
- The relative weight to allot to each task, each content category, and each skill being assessed should be specified. Typically, the weights reflect the importance that content specialists place on the particular kinds of knowledge or skills that the assessment is designed to measure. In some cases, time on task or importance for the intended use might be major considerations of greater weight than proportion of time spent.

In 'writing of scoring specifications':

- This report identifies multiple approaches to response scoring and insists they be appropriate to the tasks and purpose of the assessment as a whole. It provides details on the use and appropriateness of holistic scoring, analytic scoring, and trait scoring.
- For complex, multifaceted performance tasks such as argumentative writing or critical thinking, with clearly defined component competencies and skills to be assessed, the authors emphasize the use of a combination of holistic and analytic (or trait scoring) as most relevant. The discussion of such scoring is the most fully developed, and identifies design choices and considerations such as the selection and number of score categories; the importance of pilot testing sample tasks; and the importance of calling out the specific criteria to be considered used in formulating rubrics in a scoring guide, including performance attributes, features counts, and quality markers.
  - In general, one should use as many score categories as scorers can consistently and meaningfully differentiate.
  - The number of appropriate score categories varies according to the purpose of the assessment, the demands of the task, the scoring criteria, and the number of clear distinctions that can be made among the responses.

Note. Summarised from Baldwin, Fowles and Livingston (2005).

The focus on CR assessment in academia and industry continues to improve CR item functioning, CR item scoring, and performance assessment used in educational assessment. McClellan (2010, p. 2) is another constructed response scoring primer that framed the challenge explicitly: 'In order to have consistent and reliable CR scoring, each rater must understand and apply the scoring rubric to the examinee responses in the same way every time.' Both of these guidelines reflected the emergence of a series of practices to monitor scoring quality as it happens toward achieving this goal. McClellan called for exemplars and benchmarking, back-scoring, double-scoring, and 'trend-scoring checks' (p. 4). Exemplars for benchmarking or illustrating each quality level or score point on each measured trait or holistic score echo earlier developments in CR scoring. Within-year interrater agreement is actively measured; all papers should be scored by two graders, with discrepant results scored by a third grader. The exemplar responses can also be used to calibrate newly trained scorers or retrain scorers who show inconsistency or drift in their scoring work. Back-scoring reflects the practice of dedicating resources to rescoring some level of items from all scorers over time but includes scoring inspired by increases in discrepancy rates for particular scorers or scorers whose score distributions deviate from norms. Trend-scoring checks focus on the durability of items over time, ensuring that the items continue to perform the same way; they also pick up whether students' interpretation of an item or graders' decisions on scoring have shifted for reasons unrelated to the item itself.

Finally, as further evidence of the degree to which these processes and procedures for holistic scoring for CR items reflect a dominant mainstream assessment approach, I note that they have also been standardised and adopted as standards by the National Center for Educational Statistics (2008) at the US Department of Education. The US Department of Education has published voluminous information on CR scoring for a broad range of national standardised educational assessments that address anchor papers, practice sets and qualifying sets for scorers, and procedures for training (trainers, supervisors, and scorers) and score monitoring (with back-scoring, calibration, trend-scoring checks, and within-year interrater agreement checks).

Most of these practices for CR scoring on standardised tests have been widely adopted over the last decade. Many of the tasks—calibration exercises, double-scoring, blind-scoring, discrepant score recognition and referee scoring, scorer drift monitoring, and statistical monitoring and validation of both items and raters—are generally built into most modern distributed online scoring platforms that automate these procedures, including the dynamic and ongoing calibration operations. Economies result from allowing a fully centralised and controlled process to occur in a geographically distributed fashion. Scorer training, scorer calibration, and score monitoring routinely occur in standardised educational assessment at all levels, even as testing itself is moving in much of the developed world from paper and pencil to online activity. These benefits further contribute to the increasing consistency of human scoring for CR items with generic rubrics and category/holistic type scores.

2.2 CT and AW Scoring

2.2.1 CT assessment

It has long been argued that MCQs address decontextualised knowledge and do not access the complex, higher order cognitive skills that are important for 21st-century jobs (Gulikers, Bastiaens and Kirschner, 2004). Some hold that the development of CT is the highest objective of science education (Adey and Shayer, 1994; Bailin, 2002; Siegel, 1988). CT includes the ability to make valid inferences and logical deductions, analyse probabilities, recognise relationships, make predictions, and solve problems (Halpern, 2014; Pascarella and Terenzini, 2005). Proficiency in CT is associated with success in undergraduate education, improved life outcomes and decision-making, and a more active and engaged citizenship (Halpern, 2014). It is not surprising, then, that various stakeholders in education have long advocated a focus on CT skills in higher education (Association of American Colleges and Universities, 2005; Facione, 1990; Kuhn, 1999).

Interest in CT development and in CT as an important 21st-century skill has led to a greater need for CT assessment (Facione, 1990; Halpern, 2010; Lin, 2014). In recent years, general agreement has emerged about what CT is and how it can be recognised, described, and evaluated. The common view is that CT skills represent a horizontal or broad set of skills; as Butterworth and Thwaites (2013, p. 3) explained, ‘Critical thinking and problem solving are very broad skills, not bodies of knowledge to be

learned and repeated'. Assessments of CT skills such as the Collegiate Learning Exam Plus (CLA+) from the Council for Aid to Education (no date) or Cambridge Assessment's (no date) Thinking Skills Assessment typically include a significant writing task that requires students to make a claim, support it with evidence and reasoning, and address counterarguments.

Liu, Frankel and Roohr (2014, p. 3, Table 1; reproduced in part in Table 3-1) provided a comprehensive review of frameworks for defining CT and an examination of eight separate CT assessments. The tests cover a wide range of formats and organisations, with different tests stressing different themes, ideas, or aspects of thinking skills. They analysed the validity information from studies of these tests, which included analysis of the relationship between CT scores and other general cognitive assessments, and found moderate correlations with course grades, and with GPA, SAT, or GRE scores. They also analysed studies that considered CT scores as they correlated with negative life events, job performance, and other factors. As might be expected, they found that 'the quantity and quality of research support varied significantly among existing assessments', but noted, 'Common problems with existing assessments include insufficient evidence of distinct dimensionality, unreliable subscores, non-comparable test forms, and unclear evidence of differential validity across groups of test takers' (Liu, Frankel and Roohr, 2014, p. 7).

Liu, Frankel and Roohr (2014) also highlighted the tension in designing an assessment for CT that balances trade-offs between psychometric quality and authenticity. They framed this issue specifically in terms of 'multiple-choice items vs. constructed response items' (Liu, Frankel and Roohr, 2014, p. 4) and cited work from Lee *et al.* (2011) arguing that, in terms of testing time, MCQ items provide more information 'about what test takers know' (Liu, Frankel and Roohr, 2014, p. 8) than CR items do. They noted that an earlier study (Wainer and Thissen, 1993, cited in Liu, Frankel and Roohr, 2014) found that scoring 10 CR items cost about $30 USD, whereas the cost of scoring MCQ items to achieve the same level of reliability cost $.01 USD.[3] They also provided multiple citations of the high level of correlation between CR and MCQ item

[3] Liu, Frankel and Roohr (2014) is just a recent example from the literature and public assessment procurement data that document the cost and efficiency advantages of SR over CR items since they were introduced as a "new type of test" by Wood (1928).

approaches to measuring what is asserted to be the same constructs. But correlation does not prove that the same constructs are being measured—a key element of the arguments about SR versus CR going back to Wiggins (1990).With the best-case arguments for SR testing of CT as context, Liu, Frankel and Roohr acknowledged the inherent differences between the ability to recognise and the ability to generate. They concluded that 'in the case of critical thinking, constructed response items could be a better proxy of real-world scenarios than multiple-choice items' (Liu, Frankel and Roohr, 2014, p. 11).

Having examined a broad range of types of CT assessment and identified strengths and weakness of existing assessments and challenges to CT assessment design, Liu, Frankel and Roohr (2014, pp. 15–16) identified key elements of future CT exams:

- Evaluate evidence and its use.
- Analyse and evaluate arguments/claims.
- Understand implications and consequences (e.g., evaluate reasoning).
- Develop sound and valid arguments (e.g., demonstrate reasoning).
- Understand causation and explanation.

Rather than attempt to choose the one best framework or add to the discussion of the rationales for varying approaches to CT scoring, I have selected two common subscores or dimensions (the ability to make a logical and clear claim, and the ability to support a claim with evidence) from the most common elements of CT scoring to serve as a focus for developing CT rubrics. The common thread through these key sources of insight into CT, as illustrated in the summary findings of Liu, Frankel and Roohr's (2014) meta-analysis, is (a) the identification and use of evidence to support reasoning and (b) the ability to reason from evidence to sound conclusions. Said another way, in the meta-analysis referenced above (Liu, Frankel and Roohr, 2014, Table 3-1), three of the assessments identify claims, two identify evidence, and fully 9 of the eleven assessments cite reasoning – which requires both evidence and a claim – as explicit CT aspects they assess. Therefore, the task of scoring a response to a CT challenge must, at a minimum, be able to recognise when evidence is cited to support claims and when claims or conclusions are articulated based on such evidence. For these reasons, this study has focused on the challenge of specifying rubrics to support

scoring based on valid claims and evidence in the text of responses that align with the expectations of CT item authors.

2.2.2 AW assessment

There is a close connection between AW practice and CT skills. For example, Liu, Frankel and Rohr (2014) contains a list of 11 contemporary CT assessments, summarised as Table 3-1 and discussed in Chapter 3 of this study. This list is an update from the nine CT "standardized instruments" listed in Bers (2005, pg 17). Bers (2005) list includes earlier versions of six of the assessments that are also included in Liu, Frankel and Rohr (2014). Significantly, while both lists included "analysis" or "reasoning" in the descriptions of nearly all the assessments (7 of 9 in Bers vs. 10 of 11 in Liu, Frankel and Rohr), the latter (Liu, Frankel and Rohr, 2014) list included "argumentation" specifically for 7 of the 11 instruments described, where in (Bers, 2005), only 3 of 9 mentioned argumentation explicitly, suggesting an increased recognition of the connection between argumentation specifically and critical thinking skills. That said, a very good case is made for the CT-AW connection, and specifically for the use of AW to assess critical thinking, in Yeh (2001), and argumentation is not only listed in most of the assessments surveyed in (Liu, Frankel and Rohr, 2014), but featured in "possible assessment structural features" and the "possible tasks types" (pg 17, tables 5 and 6) as part of their key ideas for "next generation critical thinking assessment".

It is also true that the teaching of AW is closely related to teaching CT, as the two are often taught in tandem, as illustrated by the (US) National writing project[4] and Programs for CT such as the offerings of Think CERCA for K12[5] education.

Teaching AW focuses on teaching students to make claims, support claims with evidence and reasoning, address counterclaims, and so on. In this regard, the work of Toulmin (2003) informs much of the discourse on the value of AW in demonstrating and developing CT skills. The strength of the connection between the writing process and cognition was popularised by Zinsser (1988), but the strength of this connection in

[4] Examples include the (US) National Writing Project at http://www.nwp.org, and its California participants at http://writingproject.uci.edu, a structured program on Reading, Writing and Critical Thinking run at hundreds of institutions across the US.
[5] See also https://thinkcerca.com.

education, both for instructional and assessment purposes, was well developed and illuminated in the introduction to Chase (2011). Other works (Barnet, Bedau and O'Hara, 2008; Bean, 2011) explored the close connection between CT and argumentation; still others (Coirier, Andriessen and Chanquoy, 1999; Deane *et al.*, 2008) explored in detail the higher levels of cognition that underlie writing generally and, in the case of Coirier, Andriessen and Chanquoy (1999), argumentation specifically.

In most cases, AW programmes explicitly teach the subject in the context of CT. Rubrics for AW assessment (Graduate Management Admissions Council, 2016; Smarter Balanced Assessment Consortium, 2014) and CT assessment (Council for Aid to Education, 2013; Facione, 1990; Zahner, 2013) often address claims and evidence, with criteria expressed in generic terms (e.g., scoring is based in part on the response's success at making a claim and citing evidence). Such scoring leaves significant, content-specific judgements as to the right claims or the best (or even valid) evidence up to individual scorers to apply on a case-by-case basis to individual responses.

2.3 Rubrics and CT

In Myers (1980, p. 30), a work that sets out some best practices for CR scoring, the word *rubric* is parenthetically defined as 'a list of criteria' and further expanded upon as it applies to essay evaluation. Myers noted that the scoring criteria, the rubric, could be defined at the beginning of the training and scoring process if one were working with experienced readers. For inexperienced readers, Myers indicated that the best procedure is for a table leader to select exemplars or prototypes that define each score 'first, and let the rubric or list of features evolve from the discussion of the reasoning behind the scores' (p. 31). Closer to the task of scoring CT and content-centred knowledge and skills, Baldwin, Fowles and Livingston (2005, p. 1) defined rubric as 'the scoring criteria, scoring guide, rating scale and descriptors, or other framework used to evaluate responses.'

One of the earliest and most cited works on rubrics for CR, 'What's Wrong—and What's Right—With Rubrics' (Popham, 1997) established the term *rubric* in the context of performance assessment and CR as providing the basis to judge the quality

of responses. This definition held that a rubric was composed of three constituent parts: evaluative criteria, quality definitions, and a scoring strategy. The first part of that definition, evaluative criteria, distinguishes acceptable responses from unacceptable ones; criteria can be specified with different or equal weights. The quality definitions specify the way each criterion quality level is distinguished from the others. The scoring strategy specifies if and how to aggregate the quality measurements for the different criteria into a score report. This definition works for both a single reported holistic score and for score reports focused on subscores based on specific individual or groups of evaluative criteria. The definition and its three elements were precisely defined in ways that encompassed a broad array of rubric types: holistic, analytic, and combinations such as trait scoring or scores with multiple measures, each with their own quality criteria and rating scale.

In the discussion of what is right and wrong with rubrics for CR, Popham (1997) identified criteria as either too specific (Flaw 1) or excessively general (Flaw 2). This pair of concerns illustrates the difficulty of establishing scoring criteria that are at a useful and appropriate level of granularity—instructionally relevant, clear, and usable. Popham emphasised that the worst of the overly general criteria were those that attempted to distinguish between levels of quality by simply using gradations of good or bad adjectives (e.g., superior, skilled, competent, or unskilled in the ability to be measured); they provided no real guidance to how such distinctions can or should be made by the scorers (Popham, 1997, p. 4).

Although the evaluative criteria for CT rubrics cited above are not expressed or shared in item content-specific terms, there is a clear consensus on the value of evidence and the importance of claims for scoring CT or AW skills. The rubrics for the CT aspect of the CLA+ exam used in the US (Council for Aid to Education, no date), the A-level Critical Thinking Exam in the UK (OCR, 2013), the BMAT Biomedical Admissions Test (Cambridge Assessment, 2018), and Cambridge Assessment's (no date) Thinking Skills Assessment all include elements related to citing evidence and making claims.

2.4 Kinds of Rubrics for Complex Cognitive Skills

As discussed in the next section, Dawson's (2017) taxonomy for describing the elements of rubrics in assessment acknowledges that rubrics are generally classified as

holistic or analytic, and as generic or task specific. The analytic vs, holistic terminology is also used in describing rating scales, for similar reasons: analytic rubrics are expected to provide detailed criteria and may even be customized for the context and objectives of the assessment (Nordrum, Evans, and Gustafsson, 2013, pg. 922) while holistic assessments explicitly assess a single unitary construct, at times by accessing multiple facets of the construct. Between the "one score" of holistic scoring and the multiple sub-scores provided by analytic scoring, there are hybrid forms that can be described different ways. Trait scoring is popular with writing assessment, and typically specifies different aspects (or evaluative criteria) of writing as independently scored qualities of writing, each with their own quality levels and level definitions. Some hybrids of trait and analytic scoring do not attempt to combine the individual traits into a single overarching score; others explicitly do and provide their weights for the various components for the overall score.

Two hybrid scoring examples illustrate the complex possibilities employed in different high level cognitive assessments that informed the direction of this research. The first is the Critical Thinking Analytic Rubric (CTAR) at the heart of Saxton, Belanger and Becker (2012), and the other is Timmerman *et al's* 2011 "Universal Rubric for Assessing Undergraduates' Scientific Reasoning Skills" (Timmerman et al, 2011).

The CTAR study was built around a well-defined CT construct that was represented directly in the rubric developed for their proposed CT assessment. the authors devise an analytic rubric composed of six separate measures or "evaluative criteria" – interpretation, analysis, evaluation, inference, explanation and disposition – and for each of these, six quality levels. Three of these evaluative criteria had three quality levels with three quality definitions each; the other three evaluative criteria had 2 quality definitions for each of their six quality levels. Altogether, this rubric worked as six distinct "holistic" scores, with quality level definitions relying on trait-specific holistic judgements of gradations in the evaluative criteria being assesses in relative terms characterized by descriptors that range on a scale such as "unwarranted, limited, acceptable but weak, reasonable, warranted and strong, clearly justifies and explains" and "no ability, inadequate ability, uneven ability, adequate ability, clear ability and confident ability".

In short, the assessment relied on what were essentially six separate holistic measures, and particularly impressive was that with 30 and 115 students taking each of two forms of their test, they were able to achieve consistently high inter-rater reliability on all six traits or evaluative criteria with Cronbach's Alpha scores of above 0.70 (and generally well above .80). They also performed some limited blind "intra-rater testing" and achieved almost as favourable results. A close examination of the details of their scoring, however, did suggest a paucity of high scores and the sort of strong tenancy of the scorers to score using a primarily cluster of 3 or 4 midrange scores on their six point scale, with very few 1, 5 or 6 scores on most traits on most forms. This reminded me of a study (Rudner, Garcia and Welch, 2006) that showed early GMAT scores (a graduate school admissions test of the Graduate Management Admissions Council) analytical writing assessment items had a score distribution such that 87% of the candidates received scores of 3, 4 or 5 on their six-point scale.

The second study, the "Universal Rubric for Assessing Undergraduates' Scientific Reasoning Skills", while aiming for a "universal rubric", was highly focused on the discourse elements and content categories reflected in the best scientific reporting. And in this major study, which evaluated and incorporated a comprehensive set of evaluative criteria collected from professional journal referee guidelines and other academic sources, the robust and comprehensive rubric results in fifteen discreet evaluative criteria in seven categories. Each evaluative criteria had a common set of four quality levels: "not addressed; novice; intermediate; and proficient". The list of evaluative criteria by category was:

1) Introduction (context; accuracy & relevance)
2) Hypotheses (testable and considered alternatives, scientific merit)
3) Methods (controls and replication; experimental design)
4) Results (data selection; data presentation; statistical analysis)
5) Discussion (conclusions based on data selected; alternative explanations; limitations of design; Significance of research)
6) Primary use of Literature
7) Writing Quality

Most quality levels had 3 or 4 quality definitions. Reliability of individual measurements was generally good (between 0.67 and 0.85 with some stronger and

weaker measures), and the overall inter-rater reliability of a combine score that summed the traits was strong (0.85).

Based on these two ambitious hybrid rubric models, and the significant complexity of the CT construct illustrated by the 32 variously described aspects of CT identified in various CT assessments and listed in Figure 3-2 in the next chapter, it was clear that a CT rubric grounded in item-specific and content-centric definitions would require a structured set of evaluative criteria, quality levels and quality definitions that would be central to an approach designed to provide feedback by connecting the rubric elements to the item response content. Whether such a rubric is best described, in Dawson's terminology, as simply a task / activity-specific analytic rubric, or a new kind of structured rubric, the task of defining a rubric design framework would need to at least accommodate these sorts of challenges and provide a straightforward mechanism for representing the details of the scoring criteria and strategy.

2.5 Rubric Elements and Terminology

For most of the last three decades, terminology and definitions in discussions about CR scoring and rubrics have been fluid and inconsistent. Fortunately, Dawson (2017, p. 348) published a paper 'to provide a language to discuss rubrics. Rather than seek a homogenous definition for the term "rubric", it provides a framework to map out the heterogeneity of potential rubric interventions.' Dawson proposed that rubrics have at least 14 dimensions, including whether it is generic or item/task specific, and whether it is analytic or holistic. In this study I have embraced this work as a standardised set of terminology that I have adopted for describing the various aspects of rubrics themselves and for its comprehensive view of rubrics and their usage, which provides an excellent starting point for a rigorous and thorough review of possible rubrics for CR items to assess CT.

Table 2- defines Dawson's (2017) 14 dimensions or design elements, modified for simplicity and clarity from the referenced table, with the addition of notes that expand on the design element attributes. In the following chapter, each of these dimensions will be explored for relevance to the CT scoring challenges.

Table 2-1. Dawson's 14 Rubric Dimensions

Design element	Note
1. Specificity	The use of generic claims or evidence factors vs. the use of item-specific factors related to content (e.g., enumerate specific pieces of evidence vs. call for 'sufficient' evidence)
2. Secrecy	Whom the rubric is shared with and when it is released
3. Exemplars	Work samples provided to illustrate quality (at each score point or level)
4. Scoring strategy	Procedures used to arrive at marks or grades
5. Evaluative criteria	Overall attributes required of the student (e.g., what is considered unscorable or nonresponsive)
6. Quality levels	The number and type of levels of quality for each evaluative criterion
7. Quality definitions	Explanations of attributes that distinguish different levels of quality for each evaluative criterion
8. Judgement complexity	The evaluative expertise required of users of the rubric (including scorers)
9. Users and uses	Who makes use of the rubric and to what end
10. Creators	The designers of the rubric
11. Quality processes	Approaches to ensure the reliability and validity of the rubric
12. Accompanying feedback information	Comments, annotation, or other notes on student performance
13. Presentation	How the information in the rubric is displayed

Design element	Note
14. Explanation	Instructions or other additional information provided to users

Note. From Dawson (2017, p. 357, Table 1).

2.6 Summary and Relevance

This chapter has provided context for my research, which focuses on the use of CR items in CT assessment, by exploring existing standards and approaches to CR scoring and the challenges inherent in these approaches (e.g. the tension between reliability and useful feedback); by elaborating existing approaches to CT and AW scoring, and associated concerns of CT and AW assessment and scoring in particular; and by reviewing a comprehensive framework for discussing rubrics and their various aspects. Using the terminology from Dawson's (2017) framework and focusing on the CT and AW scoring challenges, the next chapter introduces a model of scorer cognition that ties elements of rubric design to the goals of improved scoring with feedback and reliability. The RDF, as articulated in the next chapter, is proposed as a way to guide the development of CT rubrics for CT by providing criteria for the various dimensions on which rubrics are constructed. With rubrics so constructed, and when applied with a clearly articulated scoring strategy, my goal is an improvement in both feedback and reliability for CT assessments in practical contexts.

Chapter 3 Development of an RDF

3.1 A Theoretical Model of Scoring

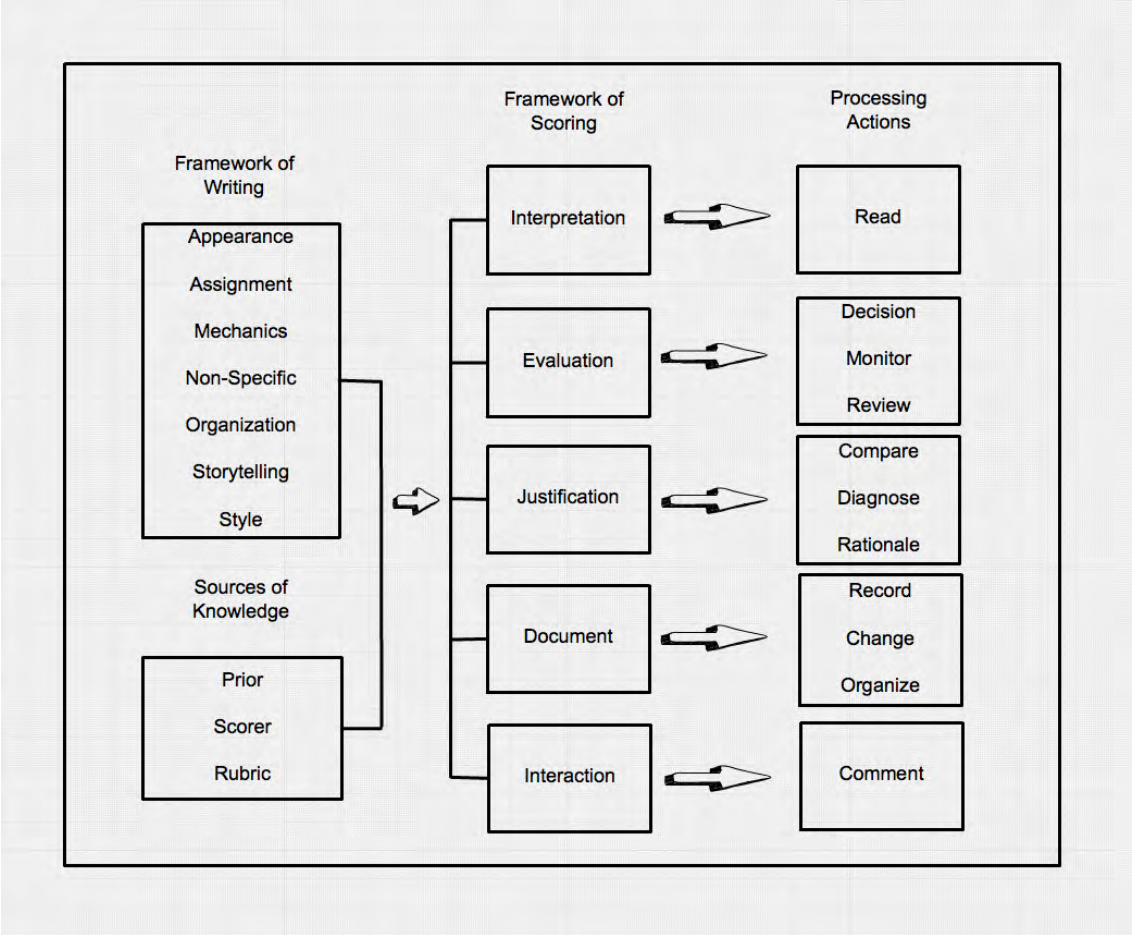
There is a broad consensus that CR assessment, compared to multiple-choice exams, is more authentic when it makes cognitive demands in a context that reflects closely the knowledge and application for which they are being evaluated (Frey, Schmitt and Allen, 2012). It can assess higher order cognition (Liu, Frankel and Roohr, 2014) and also evaluate the specific content of a response (Zhang, 2013). Content-specific rubrics are necessary if an assessment is to drive actionable feedback, as scoring that connects response elements to rubric requirements can be used to focus student and instructor alike on particular gaps in understanding and pedagogy to address specific issues.

Starting with the idea that better scoring could be informed by a thoughtful approach to how scoring actually works, I investigated scorer cognition models and found a model that described human scoring of performance tasks in terms of key data structures and key processes that used those data structures (Wolfe, 1997). The purpose of this study was to better articulate the relationship between scorer cognition and scoring accuracy. Although Wolfe (1997) was particularly looking at the possibility of improving human scoring with better scorer training that reflected this model of cognition, I recognised that this model could be useful in improving outcomes of CR assessment by using it as a basis for expressing rubrics that spoke directly to this underlying model of scoring. The remainder of this chapter develops an RDF specific to scoring CT and AW challenges as a subset of possible CR items by combining this idea with the particular goals and objectives of CT scoring and informing that approach to rubrics with the elements of rubric design as described by Dawson (2017).

From Wolfe (1997), I focused on the scoring cognition model composed of two components: knowledge representation and processing actions. An overview of this model is shown in Figure 3-1. The knowledge structures on which Wolfe's model operates translate directly as the written response and the item author's rubric. The key 'framework of scoring' shown in the middle column of Figure 3-1 identifies the primary work of scoring as actions—interpretation, evaluation, justification, and documentation—that need to be guided by the rubric. In this view, scoring is the

process by which the rubric is applied to the item response by being interpreted and evaluated against the response content. Results of this process should therefore include detailed associations between the scoring decisions called for by the rubric and specific parts of the response elements being scored. If the rubrics can lead to the establishment, recording, and reporting of these associations, then they can be included in score reports to document the scoring results and provide explicit justification for the score.

Figure 3-1. Expanded Model of Scorer Cognition



Note. Adapted from Wolfe (1997, p. 25)

Furthermore, when an item-specific rubric is being applied during this scoring process to an item response, following this model's actions—of interpretation, evaluation, justification and documentation—the scorer is clearly obligated to annotate an element of the response where an element of the rubric has been satisfied. Such annotations, if

efficiently made and correctly captured in student feedback, seem an excellent source of detailed and nuanced feedback that is one of the goals of the design framework.

3.2 CT and AW Scoring Goals and Objectives

The focus of this research is an RDF for CT assessment with a goal of useful feedback, nuanced and detailed scoring, explicit rationales for scoring outcomes, and better IRR. This section describes how these goals and the results of many studies of rubrics and their utility, effectiveness, and other effects on teaching and learning collectively inform the development of my RDF.

Jonsson and Svingby (2007) undertook a meta-analysis of 75 studies of rubrics for performance assessment to determine whether specific benefits from scoring rubrics (more reliable scoring, more valid judgements of performance quality, and greater promotion of learning) could be discerned. In the era covered by their study, even the word *rubric* was considered confusing or no more than a simple set of scoring rules (Hafner and Hafner, 2003, as cited in Jonsson and Svingby, 2007, p. 131). Rubrics were generally either holistic or analytic, when this meant either a single overall score or one score for each dimension under measurement. Jonsson and Svingby's primary findings were as follows:

- Rubrics enhance reliability of scoring for performance assessment; both exemplars (of quality levels) and scorer training in the use of the rubric are helpful.
- Rubrics (generic, holistic, or trait-level) do not enhance judgement validity but using a comprehensive framework of validity to evaluate the rubric facilitates valid assessment.
- Rubrics 'seem to have the potential of promoting learning and/or improving instruction. The main reason for this . . . is that rubrics make expectations and criteria explicit which also facilitates feedback and self-assessment' (Jonsson and Svingby, 2007, p. 141).

The implications of each of these findings for my RDF are described in the paragraphs that follow.

3.2.1 Rubric design can enhance scoring reliability

Jonsson and Svingby (2007, p. 131) were investigating evidence for

> the widely cited effect of rubric use . . . [being the source of] the increased consistency of judgment when assessing performance and authentic tasks. Rubrics are assumed to enhance the consistency of scoring across students, assignments, as well as between different raters.

They did find that IRR and scoring consistency were improved by the use of benchmarks, anchor papers, or other exemplars to distinguish quality levels. Further, rater training improved agreement, and 'topic-specific rubrics are likely to produce more generalisable and dependable scores than generic rubrics' (Jonsson and Svingby, 2007, p. 135).

These insights have direct implications for factors that could contribute to effective rubrics, which share these goals of rater consistency and reliability. The use of exemplars, which provide concrete examples that tell 'both instructor and student what is considered important and what to look for when assessing' (Jonsson and Svingby, 2007, p. 131), adds to the information about quality levels in a way that has already been translated into a form directly relevant to the representation of the student's work. This suggests that rubrics that provide specific guidance for the work being measured, not just in a generic or topic-level relevant descriptive form but content-specific response expectation as guidance for quality discrimination, could provide even further improvements in scoring reliability.

The improvement in scorer reliability from scorer training and the improvement of topic-specific rubrics over holistic ones also suggest that a rubric evaluation framework should favour more specific examples over general ones. In the same way, assessment studies have also shown that difficulty in attaining scorer consistency increases when quality levels are defined using generic qualifiers and quantifiers on a feature scale (e.g., for degree of evidentiary support, defining quality levels merely by generic qualifiers such as *minimal, some, most important, comprehensive*, etc.). Brindley (1991) and Alderson (1991) criticised use of such qualifiers as ambiguous and imprecise by providing scorers with insufficient information to consistently judge

language ability. More and better information about what signals quality levels in the context of a specific item, as unambiguously and specifically as possible, is an important aspect for my RDF to capture.

3.2.2 Rubric design can contribute to assessment validity

The meta-study by Jonsson and Svingby (2007) concluded that rubrics by themselves, based on their structure and content, could not make scoring more valid than scoring without a rubric. They explained that a rubric's real value for an assessment's validity flows from full consideration of the various facets of validity that make up the whole. A rubric's content, structure, and focus, in terms of subject matter content and necessary cognitive processes, can contribute to overall assessment validity by reflecting the thought processes and domain understanding relevant to the knowledge and skills being measured. When the scoring structure, criteria, and the rubric itself are consistent with the theory of the construct, the structural aspect of construct validity is more clearly supported.

For a CR rubric operating in the CT domain, the structure, content, and criteria should all reflect the priorities of the CT construct. Further, as my form of CT assessment is intended to support (and indeed, be) instruction, so my design goal of enabling feedback by tying rubric elements to response content provides direct support to the consequential aspect of validity. My RDF must explicitly address the importance of the intended use of the assessment as expressed in the rubric and ensure that value implications and consequential validity are addressed as part of rubric design.

Most of the studies discussed in Jonsson and Svingby's (2007) meta-analysis that addressed validity dealt with external validity. External validity is a near-universal element of any valid assessment; it would apply equally well to a CT assessment built around CR items and to my CT-focused RDF.

3.2.3 Rubrics can promote learning and help inform instruction

Jonsson and Svingby (2007, p. 139) concluded that 'the use of rubrics promoted learning and/or improved instruction, at least as a perception of the students and teachers using them'. They found that this benefit was largely the result of making expectations and criteria explicit; it was further facilitated by the ready availability of

feedback and the potential for self-assessment. A key goal of my RDF is to enable detailed feedback based on explicit and delineated criteria and detailed expectations. This harmonious alignment of goals and benefits from rubrics generally is not by chance; it is the benefit of high-level feedback enabled by item-specific, content-centric rubrics that inspires a RDF that will enable detailed feedback based on more rigorous and purpose-designed rubrics that motivate much of this work. My goal is an RDF that can preserve full transparency in rubric structure and design and deliver detailed and nuanced feedback to enable learning, justify scores, and achieve fairness and utility. To that end, my RDF leverages the strengths of generic rubrics without compromising necessary transparency or exam efficiency and fairness. This entails rubrics that have high-level attributes to communicate to students, educators, and other stakeholders that are public and useful to guide exam use as well as study and instruction. It also requires rubrics to have attributes that are not shared in advance but are specific to the exam content, such as which specific pieces of evidence are most expected, their relative value, and the full breadth of evidence that might be relevant. These hidden rubric details will be revealed either directly or by reflection in the exam feedback, which will contribute to the already complex nature of the security around large-scale assessments but be less of a concern in formative and classroom-based settings.

3.3 Rubric Design Elements for CT Assessment

With a solid working model of how CR scoring can work in CT assessment, a review of the broad range of critical goals for CT assessment and rubric design, and lessons learned from scoring with holistic and generic rubrics, I now examine the full range of the Dawson (2017) rubric elements framework to identify those elements critical to an RDF for scoring CR items for CT assessment. In the paragraphs that follow, each design element in Dawson's (2017) taxonomy of design elements is considered for its relevance to the task of scoring CR items for CT assessment. In these sections I identify which elements and which of their aspects are most relevant to the needs of CT scoring of CR items; factors for use in my RDF are identified and discussed. These elements are then be pulled together in a proposed RDF in the section that follows.

3.3.1 Specificity

Specificity is defined as the particular object of assessment, which Dawson (2017) noted can be described in a generic or task-specific way, while acknowledging that these are best seen as points on a continuum. The specific examples cited—the work of Tierney and Simon (2004) on consistency of performance criteria across scale levels and of Timmerman *et al.* (2011) on the development of a universal rubric for scientific writing—clarify the context, which is to distinguish consideration of holistic rubrics (a single metric to characterise a complex criteria) from task-specific criteria. Although task-specific rubrics can focus assessment on specific aspects of a task (e.g., quality of the hypothesis in a scientific paper, use of evidence), the evaluation for the quality levels even on these task-specific rubrics are generic by design. That is, they do not include item-specific criteria or definitions but rather are written to apply to many possible tasks or assignments (e.g., a scientific report describing the results of a research effort).

In an RDF for CT, rubrics for complex constructs like scientific writing or CT skills should serve two distinct purposes. First, they need to clearly convey the aspects of a construct that are addressed by an assessment and provide clarity around the relative value or importance of the construct dimensions being measured. This communicates the intentions of the assessment, allowing the results to be used in an appropriate fashion, and conveys a common understanding for assessor and assessees of the scope and focus of an assessment. This is important for complex, multidimensional constructs, where an exam might choose to focus only on part of a construct. For CT in particular, where many potential evaluative criteria or scoring dimensions might exist, transparency around the specific focus of the CT construct being measured in a specific assessment is necessary to insure alignment of graders, fairness to students, and validity of feedback.

Liu, Frankel and Roohr (2014, Table 2, pp. 5–6) found that different CT assessments focus on different dimensions of the CT construct, and some assessments weigh some aspects of CT skills more heavily than others. In CR scoring in particular, CT exams should be clear on how presentation and communication skills are weighted (if at all) and the role that (for example) observing the standards and conventions of written English, or of the fluidity of argumentation as compared to the weight and importance

of using evidence well, or making/supporting the correct claim. The meaning of the performance of an English as Second Language student on a CT exam could be heavily influenced by such factors. Language issues aside, exams that attempt to recognise and measure specific CT-related skills and report on such dimensions as the ability to synthesise information from different sources, critically assess the quality of evidence, or recognise and address counterarguments should reflect these priorities and their relative importance in their rubric.

Second, the rubrics for complex constructs should provide item-specific guidance for graders. Just as task-specific rubrics improve reliability and interrater consistency (Jonsson and Svingby, 2007; Myers, 1980; National Center for Education Statistics, 2008; Timmerman *et al.*, 2010) and topic-specific rubrics improve scoring consistency (DeRemer, 1998; Marzano, 2002), an item-specific, content-centric rubric will improve IRR by standardising judgements about quality levels in terms specific to the item content directly in the rubric itself. The intention is to limit the scope of individual judgement variability (e.g., about the relative value of evidence, or what constitutes sufficient or the most important evidence as articulated in generic rubrics).

Therefore, specificity as discussed in Dawson (2017) takes on more complexity in my proposed RDF, as it requires both generalisability (to set out the high-level concerns of the rubric, including the specific dimensions of the CT construct of concern, and their relative priority) and item content level specificity (to facilitate, standardise, and make explicit how an item addresses the CT construct elements). My RDF values highly both elements of definition in the best possible rubrics. Given that topic- and task-specific rubrics improve consistency even more when illustrated in detail with exemplars or a range of examples (Jonsson and Svingby, 2007), the combination of high-level scoring direction and detailed scoring specification complicates the question of how much and when the details of a rubric can and should be shared with students and others. This is described more fully in the next element of rubric consideration: secrecy.

3.3.2 Secrecy

Rubrics that are generic and not addressed to the specifics of an actual assessment item but convey what is to be measured and (in some cases) the criteria that will be

applied to a work product or examinee performance are considered useful on a wide range of dimensions. As Jonsson and Svingby (2007) argued, the communication of rubrics for complex constructs can enhance assessment validity and support learning. Simply creating a shared understanding of what is being measured and how it is being measured supports external, social, and consequential aspects of validity (Messick, 1994, 1995) by providing shared context for the use of assessment results—heightening perceptions of fairness and supporting a perspective of assessment as learning.

As my RDF supports instructional, assessment, and learning purposes, it specifies aspects of rubric evaluation that are at cross-purposes. CR items, unlike SR items, require students to generate their own responses rather than selecting them from pre-existing choices. But if scoring criteria are provided to examinees in advance, this may jeopardise the integrity of the exam. These two aspects of a rubric—the high-level definitions and the low-level scoring details, which satisfy distinct aspects of the rubric requirements—mean that my framework has two distinct parts, each with a different valence in the area of secrecy.

The RDF requires clear specification of what aspects of the CT are being measured, how the measurements are determined, and what quality levels convey explicit criteria and implicit indicators of scale and granularity. This aspect of the rubric specification broadly communicates the purpose, intent, and function of the assessment, fostering validity, fairness, and educational goals that are broad and general.

At the same time, the RDF calls for rubric elements that are detailed, content centred, and item specific to enable feedback, more reliable scoring, and the ‘assessment as education’ that comprehensive feedback can create. Such detailed rubric content, if known in advance to the testing population, would nullify the effectiveness of the assessment instrument. With comprehensive feedback and information from an item author about the quality and correctness of responses available prior to the assessment experience, learners would benefit from the material in ways that would necessarily affect their response.

Thus, rubrics should include two components. One is a clear and shared articulation of the aspects of the CT construct that are being measured, including a description of the nature and scale along which the dimensions will be measured. The other is a secret set of item-specific rubric information used by graders and authored by the item writer. This information is the basis for detailed scoring decisions and the source of material used to provide comprehensive feedback to those being assessed.

A further implication of this RDF is that assessment providers could construct different assessments for different purposes. Assessments for sorting, placement, competency qualification, or other purposes that do not include providing a learning experience or maximising what a student can learn from an exam might choose a more circumscribed score report, keeping quality indicators and potential feedback to examinees secret to allow continued use of a CR assessment of CT.

In this study my priority is for assessment as learning, including the many formative assessment scenarios in which a comprehensive score report fully realises the learning potential of the assessment experience for teachers and students alike. At least one author has found evidence that too much specificity in evaluation criteria for performance assessment can have unintended and negative consequences. Torrance (2007) warned that criteria compliance could become an instrumental, box-checking exercise that replaced genuine learning from assessment. One way to prevent this, suggested by the work of Lazer *et al.* (2010), is to include a broad range of items and assessments with variability in their design and execution, focusing on different aspects of a single more complex construct. Reliable, learning-enabled assessment of complex competencies, with tasks that practically and cognitively reflect the knowledge and skills being measured in real-world contexts, has been widely adopted in postsecondary vocational education, suggesting increased recognition of value in education-as-learning (Torrance, 2007).

3.3.3 Exemplars

Exemplars or work samples that are provided to illustrate quality levels are helpful in improving reliability in assessment of complex cognitive capabilities (Tierney and Simon, 2004). The value of such anchor papers or range-finders has been a common element of CR scoring for decades (Baldwin, Fowles and Livingston, 2005;

McClellan, 2010; Myers, 1980). Examples of how judgements across a range of as many as eight factors are collapsed into a single holistic score or trait quality measure are designed to ensure consistency of measurement among raters. When CR was first tried in large-scale assessment, the social work of collaboration (where groups of raters would review the anchor papers together and discuss the distinguishing aspects of scores at each quality level) was seen as instrumental in achieving acceptable levels of IRR (Myers, 1980).

Cumming, Kantor and Powers (2002, p. 68), noted challenges to holistic scoring:

> Holistic rating scales can conflate many of the complex traits and variables that human judges . . . perceive . . . into a few simple scale points, rendering the meaning or significance of the judges' assessments in a form that many feel is either superficial or difficult to interpret.

The conflation of such scoring information into a point on a scale can have the effect that Bejar (2017, p. 573) summarised succinctly: 'There is a price for the increased interreader agreement made possible by holistic scoring, namely that we cannot document the mental process that scorers are using to arrive at a score'. It is this lack of information, the hidden working of the scoring process itself, that removes useful and nuanced feedback from what is made available to the student.

My RDF must ensure that the traceability between score outcomes and response text interpreted by the rater is captured and stored and persists so that feedback is possible and scoring decisions can reveal insights that inform instruction or speak to the student in ways that outcomes alone cannot. The rubric evaluation criteria ensure this traceability by requiring that decision criteria relevant to rating decisions are themselves represented in the rubric (e.g., What is the correct claim? Which evidence is the most important?). This content-to-scoring association that must be captured is what requires rubrics to be more content centred, item specific, and content based and is a crucial element of my approach to scoring that addresses the concerns of this research.

It is these item-specific, content-based aspects of the rubric expression in this RDF that assessment providers will need to shield from student view prior to exam administration. After assessment, the reports and information provided could take full advantage of these data and provide detailed feedback and rationales for grading decisions, with clear indications of what response elements could be improved.

3.3.4 Scoring strategy

The scoring strategy, as described by Dawson (2017), is taken from Popham's (1997) three essential features: scoring strategy, evaluative criteria, and quality definitions. In this framing, the scoring strategy is the overarching mechanism that defines the score, built up on the basis of the evaluative criteria and associated quality definitions.

The multidimensional nature and differing views and interpretations of the CT construct are reflected in the variety of dimensions and subscores reviewed in Liu, Frankel and Rohr (2014).

Table 3-1 summarises 10 assessments from Liu, Frankel and Rohr to highlight the number and kind of dimensions these CT exams used in their definition and reporting. The number of CT dimensions measured for these assessments varied from three to seven.

Table 3-1. Ten Critical Thinking (CT) Assessments and Their Scoring Strategies

Assessment	Vendor	Scales	Scoring strategy/structure
California Critical Thinking Disposition Inventory (Facione, Facione and Sanchez, 1994)	Insight Assessment (California Academic Press)	7	Truth-seeking, open-mindedness, analyticity, systematicity, confidence in reasoning, inquisitiveness, and maturity of judgement.
California Critical Thinking Skills Test (Facione, 1990a)	Insight Assessment (California Academic Press)	6	Returns scores on the following scales: Analysis, Evaluation, Inference, Deduction, Induction, and Overall Reasoning Skills.

Assessment	Vendor	Scales	Scoring strategy/structure
California Measure of Mental Motivation (Insight Assessment, 2013)	Insight Assessment (California Academic Press)	5	Measures and reports scores on the following areas: learning orientation, creative problem-solving, cognitive integrity, scholarly rigour, and technological orientation.
Collegiate Assessment of Academic Proficiency (CAAP) Critical Thinking (CAAP Program Management, 2012)	ACT	3	Measures students' skills in analysing elements of an argument, evaluating an argument, and extending arguments.
Collegiate Learning Assessment+ (CLA+; Zahner, 2013)	Council for Aid to Education	4 CT plus 2 writing	The CLA+ Performance Tasks measure higher order skills including analysis and problem-solving, writing effectiveness, and writing mechanics. The multiple-choice items assess scientific and quantitative reasoning, critical reading and evaluation, and critiquing an argument.
Cornell Critical Thinking Test (The Critical Thinking Co., 2014)	The Critical Thinking Co.	4	Level X is intended for students in Grades 5–12+ and measures the following skills: induction, deduction, credibility, and identification of assumptions.
		7	Level Z is intended for students in Grades 11–12+ and measures the following skills: induction, deduction, credibility, identification of assumptions, semantics, definition, and prediction in planning experiments.
Ennis–Weir Critical Thinking Essay Test (Ennis and Weir, 1985)	Midwest Publications	6	This assessment measures the following areas of the CT competence: getting the point, seeing reasons and assumptions, stating one's point, offering good reasons, seeing other possibilities, and responding appropriately to and/or avoiding argument weaknesses.
ETS Proficiency Profile Critical Thinking (ETS, 2010)	ETS	4	The CT component of this test measures a student's ability to distinguish between rhetoric and argumentation in a piece of nonfiction prose, recognise assumptions and the best hypothesis to account for information presented, infer and interpret a relationship between variables, and draw valid conclusions based on information presented.
Halpern Critical Thinking Assessment (Halpern, 2010)	Schuhfried Publishing	5	This test measures five CT subskills: verbal reasoning, argument and analysis, skills in thinking as hypothesis testing, using likelihood and uncertainty, and decision-making and problem-solving.

Assessment	Vendor	Scales	Scoring strategy/structure
Watson–Glaser Critical Thinking Appraisal tool (Watson and Glaser, 2008a, 2008b)	Pearson	5	Composed of five tests: inference, recognition of assumptions, deduction, interpretation, and evaluation of arguments. Although there are five tests, only the total score is reported.
Watson–Glaser II (Watson and Glaser, 2010)	Pearson	3	Measures and provides interpretable subscores for three CT abilities/skill domains: recognise assumptions, evaluate arguments, and draw conclusions.

Note. Adapted from Liu, Frankel and Rohr (2014, pp. 5–7, Table 2).

The complex nature of the CT construct motivates assessment providers to offer multiple subscales and report subscale scores. The goal is to provide detailed information about the examinees' CT skills and demonstrated competencies. Liu, Frankel and Rohr (2014) noted that significant studies showed limited support for distinctive measurements and low reliability for the subscores offered.

To support the potential of reporting useful subscores based on specific aspects of a CT challenge related to specific content in an item, the RDF should expect CT item rubrics to have a scoring strategy composed of the following elements to support the various needs reflected in existing assessments.

- Support for one or more subscales such that
  - each subscale has a distinct description of the ability or skill associated with the aspect or dimension of the CT construct that it is intended to reflect;
  - each subscale has an explicit contribution to the overall scale score expressed in terms of its relative weight (either a fraction of the overall score or relative weight as compared to the other subscores); and
  - each point on each subscale has a descriptor associated with a distinct level of quality for that dimension.
- A single, overall CT assessment scale score based on the sum of the weighted contributions for each of the one or more defined subscales.
- Optionally, a scoring strategy may seek to transpose the overall score from the raw, sum-of-the-weighted-subscale score values to some other final

scaled score. The translation between the raw, weighted subscale scores to a final scaled score could transpose scores on a small range to a larger scale (e.g., translate a scale with 61 points to a score between 200 and 800, rounded to 10s) or from a larger, fine-grain score (e.g., 0 to 23, grouped into values between 0 and 3 based on a detailed analysis of the distribution of raw scaled scores and other internal and external evidence or factors).

Dawson's (2017) scoring strategy element includes not only how scores are defined and calculated but also the procedures, rules, and processes around how scores are obtained (human scoring, automated, or other mechanism); the number of scorers for each response or performance; and how differences in scoring judgements are adjudicated to arrive at a final score. The proposed RDF includes scoring strategy factors as an explicit part of the rubric design; all scoring procedures and processes should be defined to enhance and support the validity and reliability of the exam with its congruence to the other design elements of the CT RDF.

3.3.5 Evaluative criteria

Dawson (2017) defined the evaluative criteria as the criteria used to distinguish acceptable responses from unacceptable responses. In the case of CT assessment, the criteria for scoring are focused on the explicit definition of the elements of the construct being measured and the subscale scores defined to reflect particular levels of achievement for the constituent competencies defined. As shown in Table 3-1, no two CT assessments use exactly the same criteria or define the same subscale scores.

The RDF for CT assessment recognises the CT construct as a multidimensional construct representing a complex and interrelated set of cognitive processes. Evaluative criteria are required but could range from singular (solving problems) to multifaceted (use evidence, reasoning, and argumentation to support or challenge a claim). As seen in just 10 different CT assessments analysed in Liu, Frankel and Rohr (2014), potential CT evaluative criteria can vary greatly in kind and number. Figure 3-2 collects the various evaluative criteria for the various studies in Liu, Frankel and Rohr (2014) to illustrate the range of those assessments.

Figure 3-2 Examples of CT Evaluative Criteria

1. Inductive reasoning 2. Deductive reasoning 3. Credibility assessment 4. Assumption identification 5. Understanding semantics 6. Understanding definitions 7. Prediction in planning experiments 8. Getting the point 9. Seeing reasons 10. Seeing/recognising assumptions 11. Stating a point 12. Offering good reasons 13. Seeing other possibilities 14. Avoiding weak arguments 15. Responding appropriately	16. Distinguish rhetoric from argumentation 17. Recognise assumptions 18. Recognise the best hypothesis 19. Infer relationships between variables 20. Interpret relationships between variables 21. Draw valid conclusions 22. Verbal reasoning skills 23. Argument and analysis skills 24. Skills in thinking as hypothesis testing 25. Using likelihood and uncertainty 26. Decision-making skills 27. Problem-solving skills 28. Interpretation of evidence 29. Evaluation of arguments 30. Interpretation of arguments

Note. From 10 critical thinking (CT) assessments reviewed in Liu, Frankel and Rohr (2014).

The scoring strategy element of the RDF for CT is defined to be congruent with this evaluative criteria element and with the quality level and quality definition elements that follow. That is, rubrics have both high-level elements that define the fundamental parameters of the rubric and low-level elements that define the details of quality definitions and evaluative criteria and quality descriptors.

3.3.6 Quality levels

CT assessments in the RDF are required to define evaluative criteria, as described above. For each such evaluative criterion, a distinct set of quality levels should be defined. The number and type of quality levels defined are reflected in the scoring strategy and evaluative criteria defined above and establish the basis for the quality definitions used to distinguish the criteria on the defined scale, in terms that allow qualified scorers to consistently measure responses along a given evaluative criterion, as defined in the element that follows.

3.3.7 Quality definitions

I note here the interrelationship between scoring strategy, evaluative criteria, quality levels and qualify definitions. As Dawson (2017, p. 354) explained,

> When a rubric is shown as a table, each quality definition typically occupies one cell and represents a particular evaluative criterion at a particular quality level. Rubric users rely on these to inform judgements about quality, and they are often used as a way to explain what a particular evaluative criterion looks like at a particular level.

The central role of good quality level definitions in assessment design is acknowledged in Popham (1997) and called out explicitly in Dawson (2017) as well. Popham noted that good assessment design requires that quality definitions provide sufficiently specific guidance to enable graders to make consistent and meaningful distinctions between criteria for different quality definitions. Although this operational definition of criteria for good quality definitions is straightforward, Sadler (2009), like Dawson (2017), noted the significant interplay between the degree of specificity in the quality definitions and the importance of strength of rater expertise or judgement required for an assessment to succeed. Liu, Frankel and Rohr (2014) also reflected this trade-off and noted that, given the wide range of possible subscales for CT assessment, more research is needed if useful subscale scores for CT are the goal. They pointed to significant evidence of CT as a unitary, integrated, single skill (Liu, Frankel and Rohr, 2014, p. 13), in part due to the low degree of subscale score reliability found by researchers into CT scoring (Liu, Frankel and Rohr, 2014, p. 4).

3.3.8 Judgement complexity

Much has been written on the subject of scoring writing quality. In the discussion of judgement complexity, Dawson (2017) made multiple references to Sadler (2009), in part to contrast the use of rubric criteria described as complex with those described as analytic (defined as those that focus ‘on the structure or presence of particular information’; Dawson, 2017, p. 355). Dawson’s example is an evaluative criterion of clarity of expression with quality definitions of *bad*, *acceptable*, and *good*. Dawson (p. 355) noted that these judgements are necessarily complex and expert, while the

analytic judgements are ‘less complex, require less expertise, and … some analytic judgments on rubrics can even be made automatically by computer’.

As noted by Liu, Frankel and Rohr (2014) and cited in Section 3.3.4 above, there is a significant challenge for defining and measuring CT skills in CT assessments so that subscale scores have good internal consistency and adequate reliability. This is particularly true with CR items scored with holistic or generic rubrics. A key goal of my RDF for CT scoring rubrics is to minimise the judgement complexity or ‘evaluative expertise required of users of the rubric’ (Dawson, 2017, p. 355). My RDF for CT highlights the importance of simplifying the evaluative expertise required by graders by operationalising the judgements required to assess the defined subscale quality levels by establishing item-specific, content-based factors expected in a response to satisfy different quality levels directly in the rubric itself.

For example, rather than specify quality levels for the citation of evidence in a CT response as *minimal*, *some*, *adequate,* and *comprehensive* and leaving this judgement to the scorer, this RDF values the identification of specific elements of evidence that are expected in this case and assigns specific or relative values to different specific pieces of evidence. In this way, the quality level indicators could be minimal (two items of low importance), some (two or more minor items and one major item), adequate (both major pieces of evidence), or comprehensive (both important points and two or more minor points out of seven potential points of evidence). With quality levels specified in terms of response content, feedback—such as what evidence was missed, what incorrect evidence was cited, and so on—is easily included in scoring, for defending assessment score results, and for pinpointing specific areas for focus in future instruction.

3.3.9 Users and uses

This RDF for CT rubrics with which to score CR responses is designed to support rubrics with multiple kinds of users:

- teachers, as a way to understand an assessment and how it works and to provide feedback to students or to inform their own instruction;

- scorers and assessment providers, to provide the most reliable and valid assessment products to their customers and to perform scoring tasks; and
- students, who can be consumers of the overarching scoring strategy and evaluative criteria to focus their study and enhance their domain understanding.

In preparation for an assessment, scoring strategy level information about a rubric should be shared (along with other information about the planned assessment, its purpose, format, function, and other parameters) to clarify goals, help students prepare, and help administrators and stakeholders understand the nature and purpose of the exam. After assessment, students will further consume elements of the rubric indirectly by experiencing the feedback provided with assessment results, information embedded in scoring strategy and implementation rules, evaluative criteria, quality levels, and quality definitions that are the basis for feedback that will enhance their learning experience.

3.3.10 Creators

This RDF is designed first for use by assessment developers and item writers, with the intention that the resulting rubrics, items, and assessments will foster reliable measurement and useful feedback to students, making their assessment experience a learning experience. Generic, off-the-self rubrics might well provide the skeleton or superstructure of a useful rubric (e.g., defining the scoring strategy and evaluative criteria). But this RDF is designed to produce and highlight assessments that can provide definitive and specific feedback as part of the scoring process, with low scoring-complexity requirements and lower cognitive load on scorers than generic rubrics, supporting both formative use and education via scoring feedback.

Accordingly, assessment and item authors should use the RDF to insure a clear and logical structure to the rubric from which a score will be determined. This starts with a well-defined set of evaluative criteria. These are the primary indicators of an item or an assessments primary traits or indicators of the measured capacity: For CT, these could be related to the ability to articulate claims or identify supporting evidence, or any of the other aspects of critical thinking shown previously in Figure 3-2. As scorers must evaluate a response along the lines of each defined evaluative criteria, they

should be sufficiently distinct to contribute to the overall assessment in a meaningful way. Within each evaluative criterion, a number of distinct quality levels and quality definitions should be defined such that independent scorers and other observers can reasonably differentiate between the levels that would be associated with a given response. Quality levels with broadly defined quality definitions, or with significant overlap between quality definitions for different quality levels, will lead to inconsistent results and unreliable scoring. Making quality levels with overly narrow quality definitions, so that distinct scores do not distinguish broadly recognisable levels of mastery or understanding, can lead to unreliable scoring as different scorers make fine distinctions in different directions or make inconsistent scoring decisions when none of the overly narrow quality definitions seem to apply to a given response.

Along with clearly structured evaluative criteria and quality levels with quality definitions that are clear and distinct, item creators should minimise the subjective aspect of quality definitions by reflecting item-specific content and concepts to clarify the item author's intent in an unambiguous way. That is, the quality definition should eschew generalised terms and imprecise phrases that require the rater to apply such principles or general terms to the item content. Rather than rating levels that define *sufficient*, *some, minimal, none* or *the most important* evidence, for example, a rubric optimised by the application of the RDF would be expected to enumerate five specific and distinct levels of evidence citation. This could be done in a variety of ways, but it should be specific enough to the item and its content that any two raters would come to the same conclusion about what quality level to assign to any specific item response for each evaluative criterion (e.g., of citing evidence).

In addition to detailed quality level definitions for each evaluative criterion, the RDF requires a consistent and explicit scoring process that specifies how the assigned quality level definitions across the different evaluative criteria are combined to create an overall score for an item response—or perhaps whether a combined score is even defined. If they are to be combined into an overall score, the relative weights and the specific formula by which they are combined for total score will be part of the rubric. In addition, any scaling or other factors required to put the final raw score on some specific scale can be yet another part of the total set of scoring details captured by a well-structured rubric.

3.3.11 Quality processes

This element is defined to address the need for an assessment to ensure the reliability and validity of the rubric. The proposed RDF therefore includes a quality process factor to explicitly recognise the importance of processes for ensuring reliability and validity in the rubric. As Timmerman *et al.* (2011) demonstrated, many activities can be undertaken during rubric creation to ensure validity. These include reviewing similar rubrics for similar tasks, reviewing dimensions and quality descriptors defined in other CT assessments or related tasks (AW, performance tasks designed to reveal CT abilities), and iteratively developing the rubric based on feedback from item pilots and other stakeholder reviews.

The existence of the proposed RDF and the creation of more robust and detailed item-specific rubrics create new opportunities for monitoring scoring reliability and accuracy. The detailed associations between rubric elements and item response content (which drive scoring decisions) provide data unavailable during the application of holistic rubrics—data that could illuminate potential sources of variation in scoring. With these new rubrics in place, scoring discrepancies can be analysed on the micro-decision level. For example, if some scorers are giving credit for evidence citations that are similar to response elements that other scorers do not credit, a detailed review of the differences in scoring on a specific evidence point across an entire population of responses might reveal an ambiguity in the rubric that could be addressed with a clarified language or examples.

The scoring strategy element is another place where procedures defined for determining scores can also address reliability and validity issues. As CR assessments for CT generally include two or more graders to ensure reliability, procedures for scoring should explicitly define the procedures used to address differences in scores that might arise from two independent scorers. Johnson, Penny and Gordon (2000) reviewed some of the most common adjudication procedures adopted for double-scored CR items, including (a) combining scores from two raters (e.g., averaging the two scores) or using a third, expert scoring; (b) substituting the third score for the original two scores (and doubling the value to make it comparable to the sum of the two original scores); (c) using the third score plus the two discrepant scores, summing the three scores, and dividing by two-thirds to make the result comparable; and (d)

using the expert score and the closest of the two original scores for the final score (rather than summing the two discrepant scores).

3.3.12 Accompanying feedback

This rubric element recognises that rubrics can sometimes include feedback information, such as comments, annotations, or other notes on student performance, as part of a quality level descriptor or other rubric content. Dawson noted that rubrics sometimes include feedback within the rubric itself in his discussion of rubric users (Nordrum, Evans and Gustafsson, 2013, as cited in Dawson, 2017). This study cited by Dawsib showed that students valued exposure to the rubric's articulated feedback on holistic traits because it made the assessment more transparent and helped them to understand what was being measured.

A primary goal of the RDF is to ensure that rubrics support scoring that provides explicit feedback based on the associations that the scoring process makes between rubric elements and item response content. At the same time, public sharing of high-level elements of the scoring strategy—the evaluative criteria and the enumerated quality levels—will similarly inform students, teachers, and others of the focus and intent of an assessment, providing a degree of transparency that facilitates appropriate use.

As the RDF guided rubrics can address the presence or absence of concepts in a response, including misperceptions in addition to valid response content, feedback is possible that invokes both the top-level scoring rubric definitional elements (e.g., 'citation of evidence accounts for 60% of the score on this item') and detailed-level quality criteria ('your response did not include X, the most important evidence available, or Y and Z, which would have also provided additional support for your claim'). In short, feedback considerations are reflected in the detailed rendering of the scoring strategy, the evaluative criteria, quality levels, and quality definitions highlighted by this RDF. These quality definitions are based on expected response content, as informed by an item's particulars. Quality definitions, quality level descriptors, evaluative criteria, and scoring strategy all provide different logical places for an item author to embed context-appropriate feedback that could be leveraged when students interact with test results.

3.3.13 Presentation

This element of rubric design is concerned with the presentation of the elements of the rubric. Dawson (2017) noted this is often presented in a tabular form, with evaluative criteria down the left-hand side in the first column, column heads defining quality levels, and cells containing quality definitions. Other than an implicit suggestion that all the evaluative criteria have equal weight, this form serves as an excellent baseline to outline the basics of the scoring task and, with some augmentation, could convey the scoring strategy in full. My RDF calls for clear and explicit communication of the rubric. This study provides examples of realising these goals using a set of tabular presentations that can serve as a starting point for further discussion and evaluation. A complete, generic outline for the RDF is included as Appendix J of this study; the rubric structure and content is explicated in Sections 3.4.1 to 3.4.10.

The presentation of the complete scoring strategy needs to communicate the relative value or importance of the different evaluative criteria and provide a mechanism to unambiguously communicate the sort of content that the rubric would expect to find in a CT response at the various quality levels defined for each evaluative criterion. These definitions, in the adopted RDF element taxonomy, are the quality definitions defined above.

Detailed, content-centric rubric quality definitions for CT are not suited to the typical holistic scoring grid for a few subjective terms with short phrases on a sliding scale in each row. CT quality definitions are likely to have a higher-level component that summarises the nature of the quality level for a specific evaluative criterion that might fit into a small space. But the CT quality definitions are likely to also include by reference a detailed set of information in its own table. This additional table would describe the expectations for content in a response that would satisfy the specified quality level, perhaps with alternative ways to meet the criteria for the quality level, and explicitly relate the quality definition to the overall evaluation criterion and its role in the larger rubric. The overall scoring strategy for the rubric is also communicated as described in Sections 3.4.3 to 3.4.7, laying out the method for combining the individual evaluative quality measures into a unified whole, if a single score is to be reported, as well as any procedure necessary to reflect the relative weights of different evaluative qualities and rescaling that might be done at either the

evaluative quality or overall score levels to report scores on some final scale that more usefully satisfies the goals and objectives for the specific rubric.

Finally, a number of rubric elements must be considered in the evaluation of a rubric beyond the scoring strategy, evaluative criteria, quality levels, and quality definitions. Those aspects of the rubric will result in additional tables and other means as also defined in Section 3.4.

3.3.14 Explanation

The 14th and final rubric element defined by Dawson (2017) allows for the rubric itself to contain explanatory materials. My RDF, which is focused on the importance of feedback for CT assessments using CR items, identifies real value in rubrics that include sample score reports. Examples of the nature and degree of supporting detail for scoring will be provided to exam users, along with examples that demonstrate the utility of scoring feedback that will be part of the student reporting for the rubric. As noted in the scoring strategy and other sections above, some aspects of the actual detailed scoring information and criteria, defined as detailed quality definitions, will not be predisclosed, as they could compromise the integrity of the exam. At the same time, high-level scoring strategy, a definition of each evaluative criterion, and defined quality level descriptors should be shared in the name of transparency and fairness and as a way of communicating relevance, validity, and expectations for the exam.

3.4 An RDF for CT Items

After examining Dawson's (2017) 14 rubric design elements with the goal of CT assessment using CR items as an animating principle, this section defines an RDF for such purposes on its own terms. This section defines a set of design elements that articulate a philosophy, a purpose, and criteria with which to design (or examine the design of) rubrics for CT assessment. This proposed RDF is described below as a set of 10 potential elements in three groups: core or fundamental elements of how a construct is defined and scored; procedural or process-related aspects of how scoring logic will be implemented; and supporting elements that help illustrate the application of the rubric to items. These foundational elements provide the primary criteria by which an item rubric definition can be evaluated for suitability for use in measuring CT skills with CR items. Foundational factors directly address, at a high level, the

scope of what a rubric is designed to measure, for whom, how, for what purpose, and in what form. The procedural factors primarily describe aspects of the processes specified or implied in the rubric's conception that describe how the rubric will be applied and used. The final two RDF elements are elements that can clarify a rubric: sample score reports and scored exemplar responses.

The following discussion lays out the proposed RDF elements, using Dawson's (2017) terminology as appropriate and defining new terms when necessary. As the RDF is targeted for a specific use case—CR items assessing CT—this framework defines elements corresponding to the work of rubric design or evaluation and groups related elements to focus on the factors that make these rubrics more successful. As a result, the design framework elements do not correspond one-for-one to Dawson's rubric elements. However, this framework addresses those elements comprehensively, as described in the various RDF definitions that follow, and further illustrated with a mapping between Dawson's rubric elements and my RDF elements at the end of this section (). A summary of the RDF elements for rubrics for CR items for measuring CT that are to be examined in the following sections is shown in

Figure 3-3.

Figure 3-3 Rubric Design Framework Elements for Constructed Response Items Assessing Critical Thinking

Core/foundational elements	**Scoring processes**
1. High-level rubric definition	7. Scoring processes strategy and design
2. High-level item(s) definition	8. Scoring process implementation: instructions, algorithms and QA
3. Scoring criteria and level definitions	**Supporting elements**
4. Subscale score calculation formula	9. Format and content of score reports
5. Final raw score formula	10. Exemplars (example responses)
6. Score scaling formula	

A complete form of the RDF, identifying the major elements outlined below, is presented in a series of tables in Appendix J.

3.4.1 High-level rubric definition

The high-level rubric definition for a CR assessment item for measuring CT skills includes a definition of the construct, ability, skill, or knowledge being assessed. This definition is structured to include evaluation criteria or subscores, describe how they relate to a key facet of the construct being measured, and establish their relative weight and value in the creation of an overall item response score. This definition includes a general description of how the item is intended to measure the construct, in whom, for what purpose, and how (generally) the results will be reported. The definition should include the nature of the work product being scored, the instrument used in the assessment, and the sort of score and feedback anticipated to meet the goals of the assessment.

The purpose and use of an item, with a goal weighted toward feedback and learning, toward diagnostic measurement, or as an element of a summative, comprehensive set of measures should be in accord with the structure and content of the item. Intended use should be in harmony with design elements such as depth and detail of the scoring, the nature and detail in the feedback, and the visibility of the item and its workings to various parties: assessment programme advisers, instructors, examinees, scorers, and administrators or policy makers. The use of scores or score reports should also be appropriate in the context of what is not being measured (e.g., the degree to which writing skill, language fluidity, or basic grammar, syntax, and spelling are part of what is being scored, or irrelevant, or only relevant to the extent they inhibit the communication of ideas). Score reports and scoring criteria should also be explicit to help stakeholders understand the scope of what is being measured as relevant skills.

The high-level rubric definition should also indicate the degree to which the rubric itself or its parts can or should be shared with assessment users and examinees. For CR/CT items, release of detailed feedback and quality level descriptions prior to an assessment could compromise the integrity of the assessment itself. When possible, high-level rubric descriptions should remain generic and shareable; they should not reveal details of an item's content, rubric-level definitions, or potential item feedback that would compromise the measurement to be made.

Rubrics using this RDF for CR/CT items will generally written by item or assessment authors. Assessment authors can at times repurpose existing items (passages, prompts, student instructions) by defining new rubrics for old items focused on different aspects of the CT construct. Weights for the different subscores can reflect a focus on assessing different skills that are part of the larger CT construct (e.g., addressing counterarguments, synthesising data from multiple sources, recognising implicit information, or describing analogies).

3.4.2 High-level item structure

It is often useful to include a sample item or item definition in the rubric itself, but inclusion by reference is typically sufficient to communicate the essence. Many CT items are composed primarily of (a) a passage, passages, or a set of artefacts that might include passages, information tables, diagrams, maps, charts, and so on; (b) a prompt or challenge statement directing the examinee to a specific CT challenge; and (c) instructions for the student. However, if a rubric is dependent on the item definition in a structural way—for example, it demands synthesis of information across multiple sources, or calling out different aspects of the source material that is expected to be used in specific ways (e.g., maps, charts, etc)—it is helpful to reflect these assumptions about the item definition within the rubric to ensure that this context is not lost when reviewing the rubric details.

In summary, the high-level definition of a CR item for CT or AW assessment should enumerate and define the elements of the item, which typically include:

- prompts, a challenge or question or other demand or task definition;
- passages or other artefacts, and their nature and purpose;
- instructions for students (if context is required for the prompt); and
- any other materials that are a necessary part of the item to which the rubric will be applied.

A sample item that is illustrative but uses different content may also be useful if the structure or composition of the item is unusual in any way.

3.4.3 Scoring criteria and level definitions

This element expands upon the high-level item description by describing in detail how the scoring will be done (by enumerating the key constructs and respective evaluation criteria or subscores) and assigns quantitative weights to each of these dimensions. This element defines, for each subscore, the quality levels used to characterise the response. Each quality level has a quantitative value or point score associated with it, as well as a definition of the quality level. Quality level definitions are used by scorers to determine what quality level should be assigned to a specific response. Characteristics that help scorers distinguish between quality levels are important to define, and CR for CT scoring rubrics are most robust when specific content—the expression of ideas, reasoning, evidence, or conclusions—can be used to distinguish between quality levels.

Quality level definitions may also include quality descriptor information that can help clarify the meaning of the quality level. Quality levels and quality descriptions are often organised and displayed in a table, with quality levels in columns and evaluation criteria in rows; the cells of this matrix contain descriptions of the definition and criteria for a given evaluative quality at a specific quality level. Such tables are what most people think of as a scoring rubric; putting ‘rubric scoring table’ into an Internet search engine will generate millions of results and countless images of such tables.

3.4.4 Subscale score calculation formula

CR for CT rubrics using this framework rely on the presence or absence of specific content in responses to help scorers determine the most appropriate quality level to assign to a given subscore during scoring. In many cases, one or more concepts contribute to a specific subscore quality level, and scorers seek a variety of information in the response to determine if the response demonstrates understanding and nuance. These specific concepts in the text, or target response elements, may indicate some aspect of the response relevant to scoring for a particular subscore. Finding such content in combination with other content is part of the scoring processes. To help scorers navigate the application of the rubric to the item response content itself, subscale scoring criteria and definitions include indications or rules for how to score combinations of relevant subscore content. For example:

- Evidence that the flower was the same as the one in her home country (e.g., unmistakable scent, colour of flower, shape of leaves: max 1 point for any of these)
- Evidence that the flower was different from the one in her home country (e.g., not as pretty, texture of leaves, ability to survive cold; up to 2 points, one each, for any of these)
- Other bits of evidence relevant to the challenge
- No more than 6 total points for evidence

A more complete example of a subscore calculation formula, which is integrated with the quality level and quality definitions, is shown in Figure 3-4.

Figure 3-4 - Example of Subscore Specification

Evaluation Criteria: Use of evidence to support a claim; ‘Evidence’.

Subscore formula and quality level definitions:

Supporting evidence in any of the six categories listed below will count as one point toward the evidence score. There will be at most one point for each of the six categories specified; the total subscore value will be the simple sum of these six possible evidence points, between 0 and 6.

#	Value	Award point per category for evidence of each kind.
1	1	The winter hibiscus in the new place is different from the hibiscus back home. a. ‘It’s not a real one. Not like the kind we had before’.
2	1	Winter hibiscus not as pretty. a. Winter hibiscus is different—not as pretty, flower less beautiful than the hibiscus they knew before.
3	1	Winter hibiscus is strong enough to survive the cold/winter. a. Winter hibiscus is different—stronger/more tolerant of cold/winter than the hibiscus they knew before.
4	1	Saeng’s mother has begun to adapt to the new environment. a. Acclimation to the cold; persevere to provide continuity for her child.
5	1	Saeng has begun to adapt to the new environment; Saeng recognises survival requires determination and work, even change. a. Her mother had said survival is ‘what matters’. b. Determination to succeed, do what is necessary in the new place.
6	1	The winter hibiscus is in some ways the same as the hibiscus back home. a. Petals, blossoms, stamen colour/texture as before. b. Examining the flower met expectations (feel: cool and smooth), etc. c. Hence it has adapted/changed to accommodate the circumstances.

Furthermore, subscore scoring detail may contain additional rules and scoring definitions such that a varying number of points could be assigned to full or partial expression of a claim or argument. The presence or absence of a claim, with more points awarded for a more robust and complete formulation and fewer points for a partial articulation of a more general and valid inference, could be expressed in narrative form by defining how varying degrees of completeness could be expressed.

- Full credit (16 points) for recognising that both the immigrant young woman and the winter hibiscus had to struggle to adapt and survive in a new place, and the parallel between them
- Major credit (12 points) for recognising that either the immigrant or the winter hibiscus had to adapt and noting the struggle, determination, or challenges faced by the other
- Some credit (8 points) for noting that either the immigrant or the winter hibiscus had to adapt to survive and that this entailed growth and change
- Minor credit (4 points) for noting the underlying themes of struggle, growth, survival, or adaptation in some way
- Zero points for failing to identify the underlying analogy or any part of it as central to the story

The rubric subscore calculation must be sufficient to describe the response content that will satisfy specific evaluative criteria and quality level definitions and may be described in whatever way that provides scorers with clear guidance. These formulas can vary from simple additive examples with maximums for different subscores to complex formulas or rules that are rooted in the domain in question and the item's particular content. Scoring rules could even specify negative point values for misconceptions found in a response, common errors anticipated with the response such as invalid conclusions, missing or invalid logic and reasoning, or assigning importance to irrelevant factors.

3.4.5 Final raw score formula

The final raw score formula specifies how subscore values determined during the scoring process are combined to create a final score based on the individual subscore calculations. As with subscore calculations, the final score that combines the individual subscores may be as simple as an algebraic sum of subscores, or it may require further process based on item design and rubric design factors. For example, the final raw score calculations can be a more complex function of subscores, weighing different subscores differently or applying logic such as overriding a raw score total based on subscore thresholds. This is common in mastery exams, where minimum score thresholds may be required across multiple dimensions, and in other

cases where one subscore can outweigh all others (e.g., a ‘fatal miss’ in a diagnostic radiology exam could override all other subscore values in final raw score algorithm).

An example of a simple final raw score formula is

Final Raw Score = Subscore (claim) + Subscore (evidence).

3.4.6 Score scaling formula

A rubric may include a formula that allows the result of the final raw score calculation to be transformed to a final scale for comparability or other purposes. However, for some items it is preferable to report only subscores, such as when construct representation in an assessment crosses knowledge and skill area boundaries that make combining these scores problematic and difficult to relate to a single real-world construct.

An example of a score scaling formula draws on the examples above, where subscore ranges are 0–16 for a claim factor and 0–6 for an evidence factor, resulting in a raw score range of 0–22. In this case, for comparability to other instruments that assessed the same population on a similar construct, on a 0–4 scale, the overall 0–22 score could be transformed (and validated) with a score scaling formula as shown in Table 3-2.

Table 3-2 Example of Score Scaling Formula

Raw score range	Final score	Final score descriptor
13–22	3	Strong evidence of recognising and understanding the central underlying analogy of the text
8–12	2	Some evidence of recognising and understanding the central underlying analogy of the text
1–7	1	Minimal evidence of recognising or understanding the central underlying analogy of the text
0	0	No evidence of recognition or understanding the central underlying analogy of the text

3.4.7 Score processes, strategy, and design

This RDF proposes that the way the scoring is performed, from applying the rubric to generating a final score or any reported subscore, be explicitly provided as part of the

rubric. This RDF element addresses the high-level organisation and function of the scoring process, addressing (a) how, precisely, scoring decisions are recorded; (b) how scoring data is used to produce a score and a score report; and (c) how the information will be used to provide useful and meaningful feedback.

Fundamental to the goal of this RDF is the ability to associate specific elements in an item response with specific elements in the rubric. This critical link is the basis for justifying score results and for enabling response feedback. The scoring strategy for a CR/CT item rubric, as established in this RDF, should explicitly define how the associations between rubric components and response components are captured during scoring and how these data are used in feedback and reporting to students and other stakeholders.

Related to the scoring strategy is the anticipated judgement complexity required of human scorers to apply the rubric to an item response. This RDF values items and rubrics constructed to minimise judgement complexity in scoring, which minimises the depth and breadth of domain expertise required by human scorers. This may also minimise training requirements for scorers, although monitoring and quality assurance processes and reviews are required to validate that scorer training is adequate and scoring is consistent and reliable across scorers.

This high-level scoring strategy element addresses how scoring occurs at the level of detail of identifying whether and what kind of human or automated scoring is employed; how many and what kind of scorers will score each response; and how multiple scores for an item will be consolidated into a single item score. The overall approach to defining discrepant scores and the kind of adjudication process anticipated is be defined at a strategic level. Details of the scoring processes' implementation are defined in the next RDF element.

3.4.8 Scoring process implementation

This RDF element defines how the scoring process will be carried out and focuses on the details of how scoring works, how the results are gathered and processed, and how the RDF rubric supports quality and validity by including processes to monitor, review, and validate results. Both human and automated scoring processes are

documented, describing how training, quality assurance, validation, monitoring, and adjudication of discrepant scores are performed. Human scorer instructions and training and people-based processes to monitor and maintain scoring quality are defined, including thresholds of quality used to validate scoring accuracy, reliability, and consistency. Similarly, automated scoring processes and systems, if used, may be defined and documented in terms of data and algorithms used, quality assurance and validation procedures and methods employed, and thresholds of quality used to validate scoring accuracy, reliability and consistency.

3.4.9 Format and content of score reports

This RDF element, the first of the two that correspond to Dawson's (2017) explanation elements, addresses the clarity and power of illustrative score reports that help communicate the level of detail and kind of feedback planned for score reports. While optional, an example often communicates better than a narrative the complexities that informed data visualisation techniques can demonstrate.

At the same time, automated reporting—which should be enabled by a scoring process that captures scorer associations between rubric elements and response content—is a software exercise beyond the scope of the current study. My analysis and reporting instead identify the structure and content of desired reports and demonstrate how they are enabled by the scoring process defined. In particular, the nature and detail of feedback enabled by each scenario and for each rubric are explicitly defined for each scenario.

3.4.10 Exemplars (sample item responses, scored)

This second explanatory RDF element provides for item/rubric authors to communicate via example the range of responses they anticipate and demonstrate how better and worse, poor and excellent responses might be presented and scored. While optional, such examples—like the anchor papers, exemplars, and range-finding papers used for CR scoring for decades—can help bring quality definitions to life and illustrate examples of distinguishing characteristics between quality levels that might not be readily apparent from the text of the quality definitions. While my RDF and rubric development exercise creates three robust rubrics, the collection and curation of exemplars for future use with these rubrics is outside the scope of this study.

3.5 Dawson's 14 Design Elements Mapped to the RDF

The 14 elements of Dawson's framework are fully captured by the considerations of the RDF. In some cases the narrower scope of the RDF, focused on assessing CT-related skills with CR items, means that design elements that might be highly variable in the general case may be more narrowly prescribed in the case of the RDF. The first of Dawson's elements, 'specificity', is essentially fixed in the case of the RDF to *specific* (rather than *general*) as the goal of the RDF is to assess CT skills and capabilities based on authentic challenges with a minimum of grader-specific judgement and knowledge applied to the task. It is by making rubrics item-specific that score reports can provide detailed feedback. The link between response content and rubric requirements, and the condition of required rubric elements not present, are what enables the transformation of an assessment into a learning exercise, and enables explicit construct-relevant feedback, and provides an explicit score or rating justification.

Others of Dawson's rubric design elements are addressed across a range of the RDF elements described above. For example, the question of 'secrecy' applies to different RDF elements in different ways, so it is addressed across a range of rubric definition elements. In particular, whether detailed scoring processes, level definitions and potential feedback can be shared or public knowledge without compromising an item's utility might vary depending on the specifics of a particular rubric definition, and so this variation is identified in those places.

A complete accounting of the association between Dawson's 14 rubric design elements and the proposed RDF for CR assessment of CT's 10 elements are shown in Table 3-3. Note that the right-most column in the table simply lists the RDF elements for reference.

Table 3-3 Dawson's 14 Design Elements to RDF Mapping

Dawson's rubric design element	Where in RDF	Where Dawson's element is addressed in the RDF	Reference List of RDF elements for assessing CT with CR
1. Specificity	1,3,4,7,8	All rubrics that conform to the RDF framework are task and item specific.	1. High-level rubric definition
2. Secrecy	1–10; explicit throughout	Rubrics that conform to the RDF call out what is sharable without compromising the assessment and what should be secret. Typically, most high-level, general information about a rubric such as high-level scoring strategy, evaluative criteria, and quality level identifiers may be public. Detailed quality definitions, details of the scoring strategy, and potential feedback are secret and partially disclosed in score reports.	2. High-level item structure
3. Exemplars	10	Exemplars are primarily used in training scorers; detailed scores on exemplars that may include feedback are secret prior to item use or exams as such information can compromise an assessment.	3. Scoring criteria and quality level definitions
4. Scoring strategy	7, 8	The scoring strategy and design (high level) amount to subscore weights and breadth of quality levels. These can and should be public. Complete rubric details (quality definitions, scoring calculation detail) are secret prior to exam or item use.	4. Subscale score calculation formula
5. Evaluative criteria	1, 3	Evaluative criteria are generally associated with each subscore and defined in 1; reflected to some degree in 3–7.	5. Final raw score formula
6. Quality levels	1, 3	Quality levels within each evaluative criterion are defined in 3.	6. Score scaling formula
7. Quality definitions	3, 4	Quality definitions are defined in 3 and affected by 4.	7. Score processes strategy and design

Dawson's rubric design element	Where in RDF	Where Dawson's element is addressed in the RDF	Reference List of RDF elements for assessing CT with CR
8. Judgement complexity	1, 3	Judgement complexity is explicitly manifest in the expertise required to understand and apply the scoring criteria and quality level definitions	8. Scoring process implementation
9. Users and uses	1–10; explicit throughout	Different aspects of the RDF rubric elements are intended for multiple audiences (instructor, administrators, scorer, examinee) at different points in time. Some elements of the rubric are not shared with students prior to use but are included in score results and feedback. Other aspects of the rubric—structure of subscores and weights if combined—are shared with all users and are public. Other details (detailed scoring logic) may be secret but are reflected in public information (exam feedback and score reports).	9. Format and content of score reports
10. Creators	1, 2	RDF rubrics are item specific and generally constructed by item/exam creators.	10. Exemplars
11. Quality processes	7, 8; 1–10	The entire process of applying the RDF to rubric creation is designed to situate the rubric squarely in the context of the knowledge and skills being assessed. This may be most explicitly reflected in processes designed to foster item/construct validity (1, 2) and insure scoring reliability and construct relevance (3, 4, 7–10).	

Dawson's rubric design element	Where in RDF	Where Dawson's element is addressed in the RDF	Reference List of RDF elements for assessing CT with CR
12. Feedback information	3, 4, 8	RDF rubrics can make anticipated feedback at the level of the quality definitions (e.g., feedback appropriate for item responses that earn or fail to earn credit for specific quality level identifiers); at the subscore level when specific quality level definitions are met/not met; and in the content of score reports themselves.	
13. Presentation	Samples provided	This study is experimenting with the presentation of RDF-based rubrics for the first time; the formats suggested in this report can serve as a starting point for optimising the dense information required by these rubrics for use by different audiences.	
14. Explanation	1, 2	Explanatory information concerning the rubric or the characteristics of the item(s) for which they are intended can be included in the rubric overview or item description, or elsewhere in the RDF components of the rubric description.	

Note. CR = constructed response; CT = critical thinking; RDF = rubric design framework.

Chapter 4 Methodology for the Study

4.1 Research Questions

The research questions (from section 1.4) that drive this study are:

> 1. Can a generalised and flexible RDF for scoring CT items (as compared to generic, holistic rubrics) be successfully used to define item-specific, content-centric rubrics that can guide essay graders to provide
>
> - useful feedback to students and teachers;
> - nuanced scoring that makes the exercise a learning experience;
> - explicit, defensible rationales for scoring outcomes; and
> - better interrater reliability?
>
> 2. Are there aspects of scoring with item-specific, content-centric rubrics that work well or that make scoring easier or more efficient?

The primary research question for this study is concerned with improving useful feedback from assessment, which I hope to enable by directly associated item response content with a rubric's quality level definitions. This data capture during scoring is at the heart of the ability to make score decisions and the associated rationales explicit, and thereby easier to justify or defend, provide a basis for detailed which in turn can make assessment more of a learning experience. Better IRR, rather than the lower reliability often associated with subscores, is another goal. By creating rubrics with item-specific criteria that can explicitly tie scoring decisions to relationships between response content and aspects of the rubric, all of these goals are possible.

4.2 Research Context

Having established the goal of testing a rubric design framework for CR assessment of CT skills, finding suitable data for a scoring experiment required an enumeration of requirements for the sort of task or challenge for which such a rubric would be suitable. Given the great diversity seen in CT rubrics and assessments generally, including the many examples considered in the early chapters of this thesis, and the 32

CT aspects enumerated in Figure 3-2, and my finding from reviews of CT assessments that the ability of an examinee to make a claim and support it with evidence was among the most universal of CT aspects included in the measurements, the choice of "claim plus evidence" type scoring was chosen as the main focus in the search for suitable assessment data to support this study.

4.2.1 Data requirements

For the purposes of this study, a search was made of available public, anonymised data repositories of assessment items that meet the following criteria: a) the included a challenge to be addressed with a claim and supported by evidence; b) the domain of the challenge was defined by a set of readings or other artifacts that could be processed by a student in single setting; c) that included a holistic rubric and original scoring by a (at least) a pair of graders on a well-defined scale; and d) a set of responses that numbered into the hundreds, and represented a range of scores and a population of students that were relevant to CT assessment in an academic setting. This data of course needed to be available in a complete, integrated set and ideally freely sharable so the work could be replicated, extended or otherwise interrogated by independent researchers.

4.2.2 Data sourcing

My original involvement in essay scoring for CR items began with my work at a major test publishing company, where I had an assignment to review the technology available for automated essay scoring in 2011 and 2012. My first task was to understand in detail how various forms of essays were currently scored by human graders. This encompassed a broad review of operational assessment items, rubrics, scoring procedures, validity studies and quality control processes. In particular, I was struck by the guild-like nature of an industry that created items only to find, through rigorous field testing and analyses, that at various stages items failed to survive for use in active item pools as a result of poor psychometric performance, generally meaning their item response characteristics did were not sufficiently consistent to use reliably for measurement, either in aggregate or for when used to assess specific sub-groups (e.g. they exhibited unwanted differential item functioning).

This work included two areas of focus that inform this study. One was the trade-offs experienced by item developers between the value of item rubrics with more detail and nuance and the costs, in test time and "yield" (rate at which items survive field trials to usable deployment) of developing such items. Yield here translates directly into item and assessment development costs, because as fewer items survive reliability, psychometric and differential item performance challenges, development costs for more items must be amortized over fewer usable items, resulting in higher per-item assessment costs. And the other was the challenge of scoring items on a per-trait or subscore level with reliability approaching the reliability that could be achieved with holistic scoring. This study is focused on approach to improving what can be measured without sacrificing reliability and validity.

With this objective in mind, I approached four large, highly respected assessment companies with international reach and vast stores of items and data. I had many great and lively discussions with the staff working on these and related challenges, and these organizations has procedures in place for researchers interested in such topics. Despite pursing these procedures, I found these organizations uniformly unwilling to share the detailed data I sought. Some offered limited quantities of selected response data. I also attempted to collect my own item responses through multiple channels but had limited success at obtaining the necessary quantity of motivated responses to allow for meaningful analysis.

Finally, I turned to open-source data archives available for researchers, seeking suitable assessment items with a full complement of item data, item materials, rubrics and human scores was a challenge. While some test publishers have begun to open up anonymized essay scoring data for public examination, I found the initial data releases were largely generic "quality of writing" exercise that had little in common, in terms of rubrics, with CT or AW writing tasks. I also approached academics at multiple universities involved in assessment, writing assessment, CT and AW training programs for teachers and related pursuits. I was ultimately able to source usable data from two sources, each of which are described below.

4.2.2.1 Winter Hibiscus (WH)

The first item selected for the study was from an open archive of the Automated Student Assessment Prize (ASAP) contest, funded by the Hewlett Foundation and hosted by the data science website Kaggle.com[6]. The contest was held in 2012, and part of it was designed to have researchers attempt to emulate and predict human scores for educational writing assessments by learning from a set of holistically scored essay. Among the items used in the essays scoring part of the contest was a literary analysis question, which included a short story and a prompt which asks the Examinee to explain why the author concluded the story as she did. The story itself, 'Winter Hibiscus' by Minfong Ho, is a quiet, moving account of a moment between a young immigrant girl and her mother. The story captures a brief period of interaction between a young girl, Saeng, and her mother after the daughter's failed driving test and a circuitous route home. The story speaks to the challenges they face in the context of the struggles both Saeng and her mother face in adapting to their new country. The story unfolds by intertwining a telling of Saeng's recent encounter on the way home with a 'winter hibiscus', a flower she discovers and concludes is a form of the same hibiscus found in her native Vietnam, albeit a sturdier, not-quite-as-pretty one. Her mother points out that this flower is an adaptation of the flower they knew, one that is able to survive the comparatively cold winters of its (and their) new home.

This item was double-scored with a holistic rubric, and the contest platform makes available the prompt, the instructions, the rubric, the passage and 1,700 scored responses. A close reading of the passage and the challenge suggested a response that would require a claim supported by evidence, and while the focus on the holistic rubric was writing, the rubric did require responses to make a claim and to support it with evidence using explicit and implicit content from the item's passage. Thus I found one of the eight items used in that contest was suitable for use in this study.

4.2.2.2 Harriet Tubman (HT) and Leadership

A second set of anonymised data were provided to me by a group that runs a programme on CT and AW, the California Writing Project, at the University of

[6] See https://www.kaggle.com/c/asap-aes.

California, Irvine (UCI).[7] The complete item, which includes a passage from a book about Harriet Tubman (HT), an article on leadership, instructional materials, a writing prompt, and a holistic rubric and a set of scores by two humans using the holistic rubric. The challenge was part of an AW / CT writing intervention program and came with extensive documentation of the training and an unusually specific set of writing instructions.

This prompt requires the students to select a leadership trait most responsible for HT's success, and the instructions called out specific kinds of information that could be used in the response (e.g. ways in which HT was different from her followers, and differences in which HT's reaction to life-threatening situations was different from her followers, etc.). This combination of directions and challenge gave rise to the idea of writing two rubrics for these responses to compare with the holistic rubric – one focused on the claim and evidence selected, and another focused on the degree to which the students followed the direction on incorporating the seven different kinds of narrative elements suggested. Recognizing that CT assessment could include scoring for a wide range of response aspects, attempting to use the item-specific, content-centric approach to specifying rubric elements is a direct way of assessing the ability of a more categorical, "narrative elements" evaluative quality level definitions to support scoring work the way they might for the more detailed claim and evidence quality level definitions anticipated for CT assessment.

4.2.3 WH item details

The original WH passage, WH Item prompt and WH holistic scoring rubric are included in Appendix A. The original data source provided with this item and passage included some 1,772 responses with two human scores. Items with a 0 score in this collection generally were extremely short, unscorable, or off topic. After removing items with fewer than five words or 40 characters, about one-third of the remaining items had scores of 1, another third were 2s, and the remainder were mostly 3s with a few zeros remaining.

From these item responses, 40 were selected for rubric development and another 120 were selected for use in rubric testing. The item responses were selected at random

[7] See http://writingproject.uci.edu/.

with a constraint that once the selection process was complete, each group would have within it the same proportion of item responses at each score point as were present in the original group of responses.

The only metadata provided with this anonymised data set is that all responses were from US students ranging in grade levels from Grade 7 to Grade 10.

The item response data characteristics are shown in Table 4-1 - WH Item Response Data Characteristics:

Table 4-1 - WH Item Response Data Characteristics

	Development Data (n = 40)		Test Data (n = 120)	
	# words	# sentences	# words	# sentences
Average	121	8	120	8
Median	108.5	7	117.5	8
Std Deviation	55.96	4.73	48.21	4.28
Min	25	1	22	1
Max	251	21	230	25
Score 0 count	2	5%	2	1.7%
Score 1 count	7	17.5%	39	32.5%
Score 2 count	21	52.5%	53	44.2%
Score 3 count	10	25.0%	26	21.7%

4.2.4 HT item details

The Harriet Tubman leadership item data that includes instructions (entitled “Pathway Project Reading and Writing Assignment”, a passage from a book about Harriet Tubman (HT), an article on leadership and a writing prompt with instructions are included as Appendix C of the thesis. The original holistic are included as Appendix D. The original data provided with these items included and coded for grade in school and two grades from two raters on a 1 to 6 scale, with an overall score calculated as the sum of the two individual scores which were taken from the holistic rubric.

A total of 419 item responses were provided. Each item response had two scores on a one to six holistic scale. To ensure a representative but random sample was used in this study, these items were first organised into six discrete pools based on their “first scorer” holistic score. A computer program was then used to select randomly from

each pool (based on item numbers in each pool) in proportion to their relative size to create a group of 40 item responses for the rubric development phase scoring and another group of 80 items responses for the rubric testing phase. In both cases the selections were made such that the distribution of "first scorer" scores in each pool matched the distribution of those scores in the complete set of 419 item responses.

The result of this grouping and random selection process was such that each group had within it approximately same proportion of item responses at each of the (combined) 2 through 12 total score points as were in the original data set.

The item response data characteristics are shown in Table 4-2.

Table 4-2 - HT Item Response Data Characteristics

	Development Data (n = 40)		Test Data (n = 80)	
	# words	# sentences	# words	# sentences
Average	318	21	340	23
Median	353.5	22	340.5	22
Std Deviation	157.37	10.53	211.96	15.00
Min	42	4	33	2
Max	584	47	1102	85
Average score	3.58		3.4	
Std deviation	1.57	Min = 2	1.57	Min = 2
Median score	3.5	Max = 12	3	Max = 12
grade 07 cnt	7	17.5%	20	25.0%
grade 08 cnt	5	12.5%	24	30.0%
grade 09 cnt	8	20.0%	9	11.3%
grade 10 cnt	10	25.0%	10	12.5%
grade 11 cnt	6	15.0%	10	12.5%
grade 12 cnt	4	10.0%	7	8.8%

4.3 Scoring Protocol

This study used two raters in all cases, both for data selected for use (the "holistic" scores for responses provided with the baseline artifacts) and for all scoring work done for this study using the RDF-based rubrics that rescore these items. Jonsson and Svingby (2007) note that two raters under prescribed conditions can produce acceptable levels of inter-rater agreement. They also note, referencing several studies (pg. 135) that analytic scoring is often preferable; that agreement is improved by training; and that topic-specific rubrics are like to produce more generalizable and dependable scores. As my study used raters with similar qualifications and

experience, consistent pre-scoring training and evaluation, and worked from shared instructions and scoring notes, the prescribed conditions referenced were met as further detailed below.

All the items scored in the study were scored by two of three raters; a fourth rater scored additional items as a final quality assurance check that produced addition data not part of the results analysed for this study, but which provided a promising confirmation of the results obtained.

All four of these scorers were fluent English speakers and graduates of masters programs in English or Educational Measurement, with more than 4 years of teaching experience; 3 were actively engaged in post-graduate studies in assessment. All were engaged in calibration training prior to scoring any of the data sets. Training raters is best practice for improving inter-rater reliability (Rezaei & Lovorn, 2020; Miller and Linn, 2000) and provided helpful guidance an instruction to maximize scoring accuracy and efficiency.

The training program started with a review of item materials – passages, prompts and instructions – followed by an initial review of a small number of pre-scored items, and then an assignment to score an additional pre-selected set of 20 items that represented a full range of expected scoring outcomes. Scorers kept notes as they scored each item, and the results were reviewed soon afterward. Included in this review was a discussion of the CT construct in general and the evaluative qualities and quality level definitions addressed in the specific rubric. Specific items were discussed with a focus on scoring considerations that were consistent and reflected a common application of the rubric to the responses. At the end of the training session, a final set of additional items were scored (five or fewer, depending on results to that point) to validate any performance concerns had been adequately addressed. All of the raters successfully completed the training and went on to score full item response sets. Scorer notes or questions from each training session were accumulated and used to inform subsequent scoring sessions to further solidify a common view of the rubrics and their application. These comments were also used to inform changes to the rubrics for phase 2 of each scenario. Specific examples or scorer questions are cited during the analyses of the

phase 1 results for each scenario later in this dissertation. A total of four scorers participated in the training.

During scoring raters checked in every after 20 to 40 items at which time I reviewed time-on-task, scoring notes and scoring outcomes and answered any questions raised by the scoring team. Scorers provided written questions in some cases, particularly when challenged to apply the rubrics during the first rubric development phase of each scenario. Such feedback included item specific comments for items they had scored.

In every case the same pair of scorers scored both the development items and the testing items for a given scenario; a final set of validation scoring was done after the project was completed as additional confirmation of the results. The comparable backgrounds and common scorer training and scoring procedures describe here created the conditions defined in Jonsson and Svingby (2007, pg. 136) to maximize inter-rater reliability.

The participation by the scorers in each scenario and phase was:

- Scenario 1, Phases 1 and 2: Scorer 1, Scorer 2
- Scenario 2, Phases 1 and 2: Scorer 1, Scorer 2
- Scenario 3, phases 1 and 2: Scorer 2, Scorer 3
- Validation round: Scorer 4 scored Scenario 2, phases 1 and 2.

After each rater completed their phase 2 tasks, a follow up questionnaire (see Appendix I for questions) was sent to augment the notes taken during training and scoring sessions to assess the degree to which the rubrics were perceived to support or hinder the scoring process. The feedback from scorer notes, and the questionnaire results were examined to address the secondary research question as discussed in Chapter 7. The questionnaire queried scores explicitly on questions of a) the degree to which they found the rubrics overly specific or overly general, and b) the degree to which the rubrics form of expression and specific content had an impact on the efficiency or difficulty of the scoring work itself.

While not included in the data analysed for this study, inter-rater agreement rates for Scorer 4 (the most academically advanced scorer recruited) with Scorer 2 on Scenario 1 and 2 item responses were characterized by QWK values of 0.9288 and 0.9465, respectively – the IRR perhaps benefiting from the total accumulated scorer notes.

4.4 Research Design

This chapter introduces three scenarios that provide context for the baseline examples of CR items and rubrics to be used in this study. Each of the three scenarios compares the performance of two scorers on a set of item responses using a generic, holistic rubric with the performance of two scorers on the same item responses using an item-specific, content-centric rubric defined and optimised in the context of the RDF that is the subject of the research. Items have been sourced from a public data repository of holistically scored assessment items and an ongoing research project that teaches CT and AW, and in the process collects and holistically scores responses that are available for research.

This research starts with scored items responses and in each case included the materials used in the assessment process such as the item passage(s), the student instructions, the item prompt and holistic rubric used to score the items, and the item responses and results of two scorers rating each response. Items were selected that provided sufficient structure and content to serve my purposes (e.g., they proposed a question with a prompt that was best addressed by making a claim and citing evidence; the item content supported a range of good and poor, better and worse variations of the claim it made; and for which a variety of specific elements of evidence were also available in the item materials to provide a suitable basis for scoring.)

The process of the development and testing of new rubrics for each of the three scenarios is divided into a development phase and a testing or validation phase. Initial item-specific rubrics are defined based on the RDF identified in the prior chapter, with specifics of each element of the framework addressed during the initial rubric development phase of each scenario. Trial use, scoring, and analysis then informs some adjustments to the rubrics, which are more fully tested during the second phase

of work for a larger number of items. The methodology for comparing the differential performance of the two scorers using the same rubric for the same CT task responses, and for comparing the performance of scoring using different rubrics, is also explained in this chapter and the basis for the analyses in subsequent chapters.

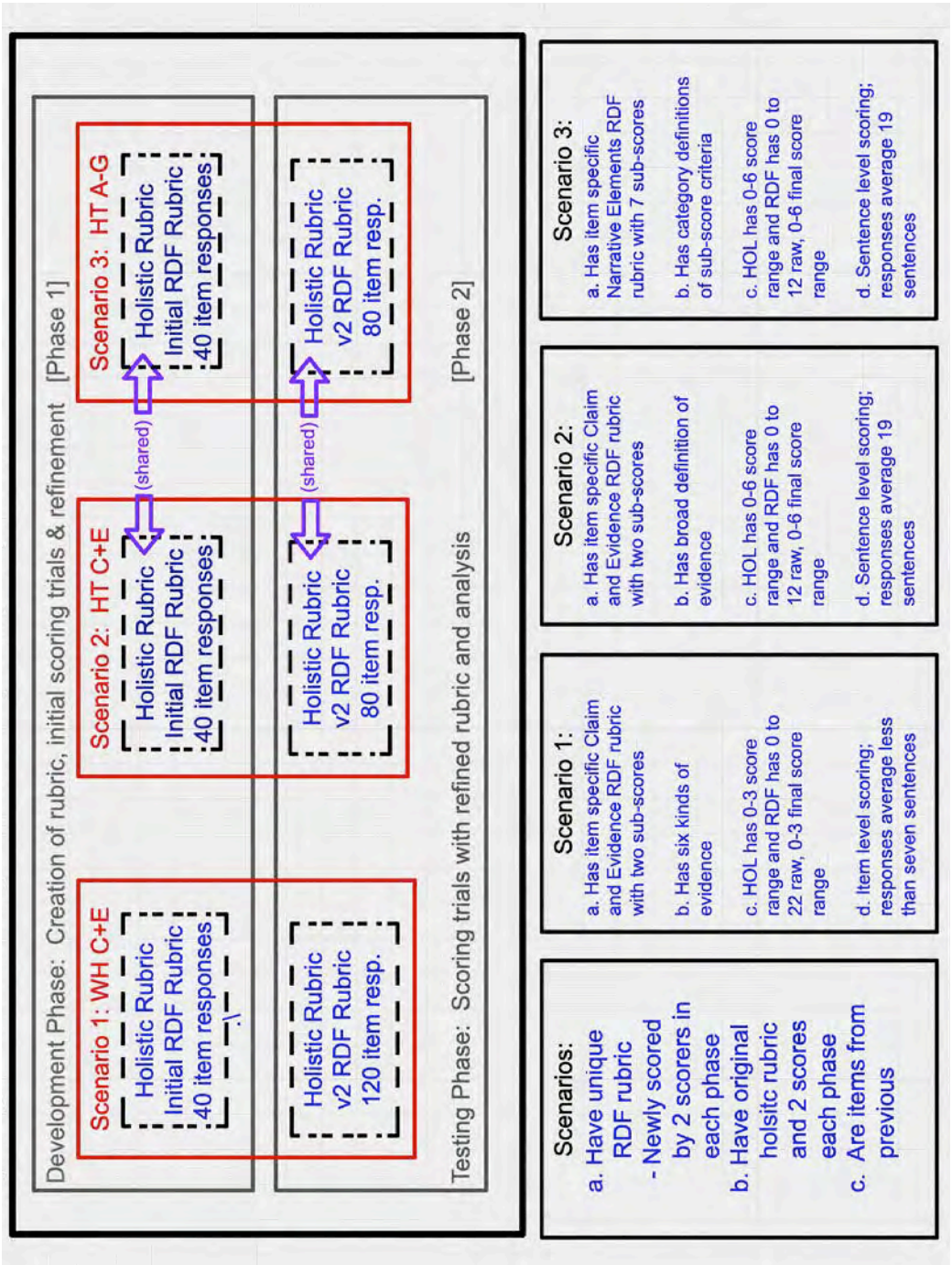
An overview of the research design is laid out in Table 4-3 below, which shows three scenarios with both development and testing phases. It is further illustrated in the schematic that follows in Figure 4-1. The general design of the RDF for CT assessment items was described in the previous chapter. This chapter describes the research design in terms of scenarios and phases. Chapter 5 develops the item-specific rubric for each scenario and describes and analyses the results of an initial scoring. This rubric development work concludes with scoring performance information from the application of the rubric, an analysis of the scoring results, and suggested refinements to the item's RDF rubric. The refined form of the RDF rubric is applied to a larger test population and analysed in Chapter 6.

Table 4-3 Three Scenarios, Two Phases

Scenario number and name	**Phase 1: Development: Score with initial RDF rubric and compare to holistic scoring**	**Phase 2: Testing: Score with revised RDF rubric and compare to holistic scoring**	**Scenario measurement focus**
1: Winter hibiscus with RDF claim + evidence rubric	40 item responses	120 item responses	Ability to articulate claim and identify associated evidence
2: Harriet Tubman with RDF claim + evidence rubric	40 item responses (shared with Scenario 3)	80 item responses	Ability to articulate claim and identify associated evidence
3: Harriet Tubman with RDF A–G narrative elements rubric	40 item responses (shared with Scenario 2)	80 item responses	Ability to follow AW instruction and narrative element guidance

Note. RDF = rubric design framework.

Figure 4-1. Study Design: Three Scenarios, Two Phases



Note. C+E = claim + evidence; HOL = holistic; RDF = rubric design framework; WH = Winter Hibiscus.

By translating the intention of the items to be scored with the new rubric, it maybe that they do not measure identical constructs. That said, the formulations are designed to sufficiently reflect the concerns and focus of the original items to provide some measure of utility in either validating this scoring approach or showing that—even when construct equivalence could be assured—the results are not sufficiently compelling to warrant further research and development.

Validity in the Messickian tradition of comprehensive validity encompassing both the narrow or domain specific construct sense (Messick, 1980, 1989), but also in the larger, broadly conceived sense of validity as demonstrated through sensitive assessment construction ethical use of results (Messick, 1994, 1995), is a primary driver for this study. CT, problem solving, and argumentation skills and knowledge are best and most convincingly demonstrated when represented and measured in their execution by the examinee, rather than by the selection of responses from which these skills might be inferred. Liu, Franel & Rohr (2014) identify the use of CR items for more authentic assessment of CT as a "major challenge" (p. 8) to designing critical thinking assessments, particularly in the context of item development costs, scoring cost and operational efficiency. Rubrics that could improve CR items in terms of measurement, scoring efficiency or IRR might also increase the success rate or yield of in CR item development, lowering overall costs and improving the cost / benefit ration for CR assessment of CT.

By identifying what part of a item response satisfies a given part of the evaluative criteria, it is possible to construct items and rubrics such that both the absence or presence of certain item response content satisfying a particular quality level definition could contribute to meaningful feedback. If different kinds of evidence are available to support a claim, and some or all are not included in the response, the deficiency can be described in terms of what evidence was not cited. And if a claim is only partially correct, it may be possible to specify what a better claim would have included. This approach informs the rubric specification task at hand and is a key to supporting the benefits that flow from expressing a rubric in this fashion.

4.5 Scenarios

The first phase of work to be performed for each of the three scenarios described below is the development of a preliminary RDF rubric based on RDF rubric elements defined in the preceding chapter. Once a new RDF rubric is defined for a given scenario, it will be used to score a set of item responses for development purposes. Next the done with this rubric will be compare to holistic scoring of the same item responses, and the inter-rater agreement for RDF scoring will be reviewed as well. Following a consideration of the analyses of the scoring results, revisions are considered and described that address the findings from the analysis. A revised version of the RDF rubric is then defined to reflect the potential improvements and will be used in the subsequent testing phase for each scenario with a new, larger set of responses scored again by two scorers. The comparative analysis of scorer performance with the improved RDF rubric is then undertaken, followed by a analysis that compares the interscorer performance of the scoring from two scorers using the holistic rubric with the results achieved with the Phase 2 RDF rubric. These analyses are combined and considered across the full range of the three scenarios with the two kinds of rubrics, with discussion, analysis and conclusions the topics of the final chapter of this report.

4.5.1 Scenario 1 – Winter Hibiscus – C+E Rubric

The first scenario is based on the Winter Hibiscus item described in sections 4.2.2.1 and 4.2.3. The original rubric was a generic, holistic writing rubric (with all the original Harriet Tubman item materials, including the source essay, the writing prompt and the rubric, in Appendix A) that asks the scorer to distinguish between ‘irrelevant or incorrect responses’ (for a 0 score) and a response that (a) ‘addresses the demands of the question’; (b) ‘addresses the demands of the question, although it may not develop all parts equally’; and (c) ‘may show some evidence that some meaning has been derived from the text’, for 3, 2, or 1 points, respectively. While there is a bit more to the rubric, it is written this way so that it can be used unchanged on numerous other questions with different source material. As the central underlying analogy is central to what the story is about, a response to the prompt that neglects this aspect of the story entirely demonstrates a limited ability on the part of the student reader to understand the material at any real depth or to get beyond the literal words on the page and to grasp the full meaning of the passage.

To interpret the scores for their depth of CT, a new rubric was constructed. The item-specific, content-centric rubric is focused on the most basic and common elements of CT skills assessment—the ability to articulate a claim or proposition, and the ability to support a claim with evidence. A new rubric stipulates that most of the points are to be awarded for the formulation of a claim that reflects the central underlying analogy of the passage, with lesser points for a subset of the full analogy, and minimal but nonzero credit for recognising the role of determination, effort, or growth in dealing with new challenges or circumstances. A lessor portion of the overall credit, about one-third, is awarded for citing evidence from the passage. The rubric identifies six kinds of evidence, each rewarded with a small credit, including specifics such as evidence that the winter hibiscus had adapted; that the winter hibiscus was related to the hibiscus they knew from their home country; or that the Saeng or her mother were adapting to their new environment (see Appendix D for the complete rubric for Scenario 1).

The detailed, item-specific rubric includes subscore definitions for the CT aspects captured by the scoring (i.e., 'Claim' and 'Evidence' scores), and included elements in its definition (a final score formula) to reflect the initial 0-to-3-point scale, with appropriate level definitions in correspondence to the original rubric's formulation.

The original data source provided with this item and passage included some 1,772 responses. Section 4.2.3 described how 40 item and 120 item response groups for rubric development and testing activities or phases for this scenario were created to represent a range of item response qualities (as measured by one of the holistic scorers) to ensure samples representative of the larger group.

All these item responses were accompanied by two (and in some cases three) human scores on the 0–3 scale using the original generic rubric. As noted above, in both Phase 1 and Phase 2 groups, item responses were selected at random from groups that were evenly spread across the original holistic score range (where the original scores were themselves roughly evenly distributed across the range of the original holistic score range)—all with the caveat that the zero-scored item group was smaller than the others and much smaller after removing extremely short responses.

4.5.2 Scenario 2 – Harriet Tubman – C+E Rubric

The second scenario is based on the Harriet Tubman item described in sections 4.2.2.2 and 4.2.4. The original prompt for this item asks the examinee to make a specific claim about 'the most important leadership characteristic' that allowed HT to be successful, based on the materials provided. The original rubric for this item is a holistic rubric that includes elements of specificity to the HT item materials but remains subject to significant interpretation when applied by scorers to specific item responses. For example, the original rubric assigns a top holistic score for a response that 'presents a thoughtful/insightful claim about the quality of leadership that was most essential in enabling Harriet to inspire the slaves' and also presents 'specific examples of several obstacles Harriet and the slaves faced and perceptively discusses how a key leadership quality helped Harriet overcome these obstacles'. In the manner of holistic rubrics, this rubric provides varying levels of description and score for lesser levels of 'insightful claims' and 'examples', but no specific guidance about how a scorer would deal with any particular combination of those attributes. But as is typical of generic rubrics, actual examples or an enumeration of what should be considered as evidence, or what evidence would apply to any of the seven traits described in the second passage, or any other specific guidance about facts or content important for use in argumentation and their relative importance, are left to the individual scorer's judgement.

Here the RDF-optimised form of the original rubric was designed to rewards a claim statement that responds to the prompt, which requires a 'claim about the quality of leadership that was most essential in enabling Harriet to inspire the slaves' and evidence to support such a claim. The original holistic rubric (Appendix D) defined a single score with six possible values for overall quality and provided between seven and 11 distinct quality level definitions for each of them—with no specific guidance about which overall score to select when a single response can be characterised by a collection of quality descriptions associated with different quality levels (e.g., a great introduction but no real claim; or where a single response both 'gives examples of obstacles Harriet and the slaves faced and thoughtfully discusses how a key leadership quality helped Harriet overcome these obstacles' indicating a score of 5, but also 'has errors in the conventions of written English, many of which interfere with the author's message', indicating a score of 2).

To compare the original scoring with an RDF-inspired rubric, a new rubric was constructed with a target score range of 1 to 12, reserving 0 for unscorable responses and those that were off topic. A new more structured rubric was developed, guided by the RDF, to score this item; a one-page summary of this new rubric is shown the first page of Appendix E. The rubric was structured to award up to 4 points for a claim that responded directly to the demands of the question (e.g., identification of a single trait from the seven traits in the secondary passage that was most responsible for HT's success) and 8 points for citations of evidence from the materials to support the trait selected, reflecting that relative balance of factors noted in the essay scores for holistic scoring.

4.5.3 Scenario 3 - Harriet Tubman – A - G Rubric

Scenario 3 uses the anonymised item response data concerning the HT item from Scenario 2, augmenting it with additional data and context from the underlying writing intervention programme. This rubric will evaluate the item responses to determine how well the students conformed to the specific instructions regarding seven specific types of content they were advised to consider for inclusion when creating their responses. These instructions can be seen in the original rubric and materials in Appendix D and are summarized in the "A- G Summary Chart" on the first page of Appendix F. The narrative elements suggested were (a) making an explicit claim for the most important trait; (b) supporting the case for why the chosen trait was critical; (c) describing how HT's response to life-threatening situations was similar to the responses of her followers; (d) describing how her response was different; (e) describing the ways in which HT was similar to her followers; (f) describing the ways she was different; and finally (g) identifying any conclusion or generalised lesson from these observations.

For the purpose of this new rubric, I associated with the seven narrative elements the letters a through g; the RDF rubric I constructed for this exercise is therefore referred to as the 'A Through G RDF Narrative Elements' rubric. The same item responses scored for Scenario 2 are scored here in Scenario 3 but with this entirely new rubric. As we are scoring for the presence of these seven kinds of narrative elements or topics, the rubric defined seven subscores named with the letters a through g using a simple set of quality level descriptions. During scoring, each response element

qualifying as reflecting the definitions of elements a–g were each awarded 1 point. One sentence could in some instances qualify in multiple categories. Most of the subscores were defined with a maximum value of 2 points, whereas two of the seven elements had a maximum score of 1 reflecting a judgement about the relative merits of the elements in the context of supporting a reflection of CT skills. Specifically, of the seven categories, the least frequently cited and perhaps least relevant to the primary task (describing what made HT an effective leader) were element E, 'how was HT similar to her followers' (as the focus was of course on what made her different); and element G, 'what lessons could be learned' from the story, which is generally tangential to the important trait identification task.

This scoring was recorded at the sentence level; scorers were told to assign the first sentence of any multisentence narrative element as the locus of where a rubric element was satisfied in order to simplify data capture and reporting. The resulting raw score range using this rubric was therefor from 0 to 12, with five 2-point maximum value subscores and two 1-point categories.

For Scenario 3, like Scenario 2, there were again 30 item responses scored for Phase 1 and 60 item responses for Phase 2, of the same average 347 word count and 19 sentence sentence count, and so on, as found in Scenario 2.

4.5.4 Scenario summary

The table below identifies the items and rubrics associated with the three scenarios and the appendix where each is located at the end of this report.

Table 4-2. Items and Rubrics Associated With Three Scenarios

Scenario	Item name	Original rubric	Original instructions	RDF rubric + instructions
1	WH	Appendix A	Appendix A	Appendix B (claim + evidence)
2	HT	Appendix C	Appendix D	Appendix E (claim + evidence)
3	HT	Appendix C	Appendix D	Appendix F (narrative elements A–G)

Note. HT = Harriet Tubman; WH = Winter Hibiscus.

4.6 Holistic Rubrics and Artefacts

While included in their entirety in the appendices, this section reviews and highlights the distinguishing characteristics of the original materials that support each of the three scenarios and serve as a starting point for understanding the research tasks that follow.

4.6.1 Scenario 1 baseline artefacts

4.6.1.1 WH item passage

Appendix A presents the WH materials for the original item. It includes a passage, the prompt (which is all the instruction that is offered or required), and the rubric used for scoring. After the first page, which provides the URL to the Kaggle.com web site where the item materials are available for download, the next three pages contain the item's passage: a short story, 'Winter Hibiscus' by Minfong Ho. The story of nearly 1,400 words conveys a moment in time for an immigrant girl struggling with the changes her recent immigration has brought about. It includes a disappointment, an encounter with a flower that gives the piece its title, and a brief interaction with her mother on returning home from an unsuccessful driving test. It is clearly and evocatively written and operates on multiple levels.

4.6.1.2 WH item prompt and rubric

The fourth and final page of Appendix A includes three headings: Prompt, Rubric Guidelines, and Adjudication Rules. The brief rubric is a typical holistic, generic CR item rubric that provides minimal guidance to scorers, requiring considerable case-by-case interpretation by the scorers. The rubric is contained in Figure 4-2.

Figure 4-2. Scenario 1 Original Holistic Rubric

Rubric Guidelines

Score 3: The response demonstrates an understanding of the complexities of the text.

- Addresses the demands of the question
- Uses expressed and implied information from the text
- Clarifies and extends understanding beyond the literal

Score 2: The response demonstrates a partial or literal understanding of the text.

- Addresses the demands of the question, although may not develop all parts equally
- Uses some expressed or implied information from the text to demonstrate understanding
- May not fully connect the support to a conclusion or assertion made about the text(s)

Score 1: The response shows evidence of a minimal understanding of the text.

- May show evidence that some meaning has been derived from the text
- May indicate a misreading of the text or the question
- May lack information or explanation to support an understanding of the text in relation to the question

Score 0: The response is completely irrelevant or incorrect, or there is no response.

4.6.2 Scenarios 2 and 3 baseline artefacts

The original HT item with its holistic rubric serves as the basis for Scenarios 2 and 3, where the new rubric scores the same item with different rubrics to assess different aspects of the same responses. The original item materials are included as Appendix C and contain several parts, some of which contribute Scenario 2 and others contribute to Scenario 3.

4.6.2.1 HT item directions

The Pathway Project Reading and Writing Assessment materials in Appendix C begin with directions and instructions that take the students through a series of exercises before writing the paper required by the assessment. These directions span the first

four pages. These initial instructions guide the students through a critical examination of the primary passage and the secondary materials on leadership, asking the student to think about the prompt they will answer. The instructions also provide questions to help the students think about the traits defined in the secondary leadership passage, the traits displayed by HT in the primary passage, and the way to organise these ideas so that they can become narrative elements in their required essay. This is important because although Scenario 2 will examine the responses from a claim and evidence perspective, Scenario 3 is intended to provide an alternate analysis of the degree to which the students' writing reflects the thinking, instruction, and narrative elements advocated during the training phase of the assessment exercise.

4.6.2.2 HT item primary passage

The primary passage for the HT item is a four-page, 2,400-word excerpt from *Harriet Tubman: Conductor on the Underground Railroad* by Ann Petry. The passage provides a glimpse of HT's work during the Civil War and describes how she made multiple journeys into the South to guide slaves to freedom. It recounts in particular one journey where she led a group of 11 enslaved people from inception and planning through the challenges and dangers along the way, to their ultimate achievement of freedom. The passage provides ample evidence of the common concerns and circumstances of HT and her followers, and of their differences in motivation, in drive and spirit, and in the ways they reacted to hardship and life-threatening situations. After the four-page story there is an additional page of vocabulary information.

4.6.2.3 HT item ancillary passage

The next two pages are devoted to an article entitled 'Seven Qualities of a Good Leader' by Barbara White. This article provides definitions for seven leadership qualities and serves as the basis from which students are to analyse HT's most important leadership trait, as demonstrated and described in the primary passage.

4.6.2.4 HT item prompt

The item prompt is a full-page document that instructs students in some detail what to write (e.g., 'write an essay in which you make a claim about ONE quality of …') and what content should be included in the body of the essay. These instructions reinforce

the directions provided earlier and clearly direct the students to make a claim and support it with evidence. They further reinforce the direction to use the narrative elements defined earlier in the assignment. As a result, this study evaluates these responses from both of these perspectives using rubrics attuned to the criteria associated with each.

4.7 Research Participants

4.7.1 Student responses

Anonymized student response data sources were described in sections 4.2.2.1 and 4.2.2.2. In short, for scenario 1 the graders are known to be from a particular school cohort that ranges across several years; for the data used in scearnios 2 and 3, the grade level cohort for the response writers is known and reported, and again, they represent a range of grade levels between 7 and 12 (in US primary education, typically 12 to 18 year olds).

4.7.2 Response graders

Essay scores from the holistic rubric were provided as part of the data sets for the items used, and this scoring was used to satisfy the holistic rubric scoring data requirements for each scenario. Graders were recruited and trained to rescore these responses for each of the three scenarios using the newly devised RDF-based rubrics that were developed during this study. The raters were recruited directly and indirectly through the National Council on Measurement in Education's[8] Graduate Student Issues Committee from graduate students in assessment, CT, and AW at multiple institutions. Four raters were selected to assist with this work. One was an associate professor with 4 years' teaching experience in English and social science and prior research experience with scoring essays and annotating text for natural language processing tasks. Another two were doctoral students in assessment with some experience teaching English and other subjects and scoring writing. The fourth scorer had both master's and bachelor's degrees in social science and 4 years' teaching experience.

[8] See http://www.ncme.org/

The protocol for selecting, training and managing the scorers during the scoring process was described in detail in section 4.3, above.

4.8 Comparative Scoring Techniques

The scoring under the varying scenarios was evaluated in multiple ways, including on the degree to which two scorers agree on their assigned scores to the same responses and on the range of scores provided by each scorer on the overall population of scores assigned. In addition, descriptive statistics are considered, such as average, median, distribution, min, max, and variance (or standard deviation) in the population of scores they each produce on the same set of item responses in aggregate, and on how closely their scores for the same item match at different overall score points/quality levels and in the face of different kinds of scoring challenges as signalled by their own comments or based on an analysis of the underlying sources of such differences generally in the data sets.

The degree of match between two sets of scores and how they compare in terms of scores assigned to the same papers was assessed in four ways:

- **Confusion matrix.** This visual representation shows in tabular form how many data points exist for each pair of measurements in a population of score pairs—scores measuring the same attribute on a scale of 1 to N displayed in an $N \times N$ table, with score values of 1 to N from Scorer 1 labelling columns along the top axis, and score values of 1 to N from Scorer 2 labelling rows down the left side of the table. The values in each cell at column x, row y, represent the number of instances or observations that received a score of x from Rater 1 and y from Rater 2. Numbers down the diagonal, from top left to bottom right, represent the instances of agreement between the raters. Score pairs more distant from this diagonal signal greater differences than score pairs nearer to it.
- **Quadratic weighted kappa (QWK)**. This number captures the degree to which one set of scores from one process or scorer agrees with another set

of scores for the same item responses. It characterises the variance between the two sets of values and provides an adjusted-for-chance measure that, by using quadratic weighting to emphasise the effect of differences that are greater in an exponential way, and provides a single scalar value to communicate something important about the degree of agreement between the two sets of numbers.

- **Agreement**. For a population of *N* data point pairs, what proportion or percentage of that population has identical values from both raters?
- **Adjacency**. For a population of *N* data point pairs, what proportion or percentage of that population has values that are identical each other or separated by a single point? Of course, with fewer points on a scale, the degree of adjacency due to chance is higher, as is the proportion of items that will necessarily be adjacent due only to chance.
- **Additional metrics.** Additional metrics are provided in the scoring data population descriptions and can comparisons, and include descriptive statistics such as average, median, and standard deviation to characterise a population of scores.

Cohen's Kappa, the unweighted Kappa variance measure, considers all non-matches as equally wrong / different; linear weight for distance measures between score points in other circumstances might be appropriate. Note also that the Landis and Koch (1977) classification system for interpretation of kappa results classifies 0.0–0.20 as slight agreement; 0.21–0.40 as fair agreement; 0.41–0.60 as moderate agreement; 0.61–0.80 as substantial agreement; and 0.81–1.00 as almost perfect agreement. These terms will be cited in discussing results in future chapters.

Chapter 5 Rubric Design Framework Development (Phase 1)

This chapter reviews the holistic rubric, the initial development of an RDF rubric, and the results of the application of each of these rubrics by two scorers to an initial set of 40 item responses for each of three scenarios defined earlier in this work. The goal is to consider the performance of the proposed RDF and RDF-based rubrics in terms of the goals for the RDF rubrics, and for the fundamental questions addressed by this research:

- Can this scoring provide useful feedback, nuanced scoring that enables learning, and defensible rationales for scoring outcomes?

- Does it support improved IRR, as compared to the holistic scores?

The goal of the analysis in this section is also to identify potential improvements that can support these goals before applying the approach to a much larger number of item responses that will be studied in greater detail.

Accordingly, the results of using the initial RDF rubric in each case are compared to the results of using the holistic rubric, both from the perspective of comparing the two sets of scoring outcomes as well as comparing the IRR performance in the two cases. As a result, adjustments are proposed to the RDF framework and each of the rubrics, which are then detailed and applied to larger item response data sets in the chapter that follows.

5.1 Scenario 1: Winter Hibiscus

5.1.1 Holistic rubric

The holistic rubric for the WH, described and shown in Figure 4-, and included with all the other original item materials shown in Appendix A, was a simple 4-point scale that assigned a 0 for unscorable or off-topic responses. Otherwise, it was to be assigned a number between 1 and 3 to reflect the rater's view that the response either (a) demonstrated at most a minimal understanding of the text and the task, and should be scored as 1 point; (b) addressed the demands of the task and went beyond the literal meaning of the text, and should be scored at 3 points; or (c) was somewhere in the middle and should be awarded 2 points.

Items were scored by two raters each using the holistic rubric, and both scores were provided with the original response samples. Some items include a third score where there is a discrepancy between the two human scores of greater than 1 point (on a 0 to 3 point scale).

5.1.2 RDF rubric

The RDF-based rubric devised for Scenario 1 addresses the 10 aspects of the rubric design framework set forth in Section 3.4.

5.1.2.1 High-level rubric definition

The WH item was selected from the ASAP competition data store on Kaggle.com for use as a CT assessment item designed to measure an examinee's ability to read and understand a literary text, to understand both explicit and implicit meaning, and to assess from evidence provided the most important elements being communicated by the passage. The item requires the student to make a claim about the passage (why it concludes as it does) and to support the claim by citing evidence from the passage. This rubric identifies item-specific, content-based indicators of quality to facilitate consistent and reliable scoring. The overall CT score is based on a combination of the quality of the claim and use of evidence, yielding a score of 1, 2, or 3 or, if unscorable/nonresponsive, 0.

Table 5-1. Scenario 1, Phase 1: High-Level Rubric Definition

1. Rubric definition	WH RDF critical thinking
(a) Construct: skill, knowledge or capability measured	Critical thinking: Understanding a literary text and the ability to make a claim and support it with evidence.
(b) Audience	Student with 8 to 10 years of schooling who can read English as expected for this grade level.
(c) How assessed	The item consists of a literary passage of about 1,500 words and a single prompt requiring the student to make a claim based on their understanding of the text and support it with evidence.
(d) How scored	A proper claim that recognises the implicit and explicit meaning of the text contributes at least 2/3 to the overall score, and partially correct answers allowed. Citation of evidence provides less then 1/3 of total credit.
(e) Security/disclosure	As the item requires the recognition of an implicit analogy, exposure of this rubric or the scoring key will make that information explicit and compromise the item's utility outside of formative contexts.
(f) Anticipated use	An exercise to gauge a student's ability to make inferences, articulate claims and cite evidence from an intermediate-level text. With robust feedback, students can learn from score reports any specific deficiencies identified in their response and how to address them. A single data point that can help gauge argumentation skill in a scenario that requires close reading.

Note. RDF = rubric design framework; WH = Winter Hibiscus.

5.1.2.2 High-level item structure

The item used in Scenario 1 is comprised of a passage and a prompt. For purposes of measuring CT skills, this scenario includes one variation that scores the response to the passage and prompt with a holistic rubric and another variation that uses this rubric as defined in this section.

Table 5-2. Scenario 1 Phase 1: High-Level Item Structure

2. Item definition	WH Item: Passage, prompt and instructions
(a) Passage	The item uses a short story, 'Winter Hibiscus', by Minfong Ho.
(b) Prompt	The item instructs the student to read the passage and address a question about the author's choice of final paragraph.
(c) Instructions	The prompt text includes instruction that the student should support their answer with details and examples from the story.

Note. WH = Winter Hibiscus.

5.1.2.3 Scoring criteria and level definitions

The RDF criteria require explicit scoring criteria (or what Dawson, 2017, called 'evaluative criteria'), level descriptors, and level definitions. The initial RDF for Scenario 1 has the following scoring criteria, level descriptors, and definitions. The primary factor for scoring purposes is reflected in the points awarded for the recognition and articulation of the central underlying analogy in the passage. It requires the student to recognise implicit connections between topics in the passage. Points for full or partial recognition of the analogy and for relevant evidence are awarded and combined into a total raw score and scaled to yield a final scaled score on a 0- to 3-point scale, comparable to the holistic scores provided by the alternate scoring for this scenario.

Table 5-3. Scenario 1, Phase 1: Scoring (Evaluative) Criteria

3. Scoring criteria	WH Item: RDF C+E rubric
(a) Claim subscore	Ability to recognise an analogy and articulate a claim
(b) Evidence subscore	Ability to articulate supporting evidence for reasoning

Note. C+E = claim + evidence; RDF = rubric design framework; WH = Winter Hibiscus.

Table 5-4. Scenario 1, Phase 1: Level Description and Quality Level Definition

3. Level descriptors	Quality level definition (C+E)
Claim subscore	Choose the score that best fits the claim in the response.
(a) Full credit/excellent claim: 16 points	(a) Full recognition of the underlying analogy
(b) Partial recognition of the underlying analogy: 12 points	(b) Recognition of the importance of adaptation for immigrants/WH
(c) Partial recognition of some of the analogy: 8 points	(c) Recognition of analogy between Saeng/immigrants and the WH
(d) Recognition of one aspect of adaptation or determination: 4 pts	(d) Recognition of the importance of growth, struggle, determination, or adaptation for survival
(e) No recognition of the central underlying analogy	(e) No recognition of the central underlying analogy in the story or any of its major aspects
Evidence subscore	One point for each
(a) 1 point	Winter hibiscus is different from the hibiscus they knew before.
(b) 1 point	Winter hibiscus's flower not as pretty as the familiar one (different this specific way).
(c) 1 point	Winter hibiscus is strong enough, able to survive the winter/cold/snow (different this specific way from the familiar version).
(d) 1 point	Did what she had to do each day to give a good life to her child. Or Saeng is changing, beginning to experience her new environment as the new normal.
(e) 1 point	Persistence and determination are important. Adaptation is important, as survival is all important. Saeng is persisting/determined to survive.
(f) 1 point	The winter hibiscus here was still recognizable as related to the version they knew before; some aspects were the same: blood-red blossoms, five petals, long stamen, yellow pollen, feel of petal were 'exactly as expected'.

Note. C+E = claim + evidence; WH = Winter Hibiscus.

5.1.2.4 Subscale score calculation formula

For the claim score, the recognition that the response identifies the underlying analogy between the adaptation of the WH to its new land and Saeng (or the immigrants) to their new home results in a full score of 16 points, sufficient by itself for a top score.

Lesser scores for recognising parts of the analogy were attempted with this first iteration of the rubric. This rubric assigned scores for the claim, 0/4/8/12 or 16, which is by itself the claim subscore. The up to six points of evidence available for recognising component elements of the analogy are summed for the evidence subscore. The evidence subscore can result in a nonzero overall score for the student, even if the analogy itself is not recognised.

5.1.2.5 Final raw score formula

The final raw score for this item is simply the sum of the claim score and the evidence score, which is a number between 0 and 22 (16 + 6).

5.1.2.6 Score scaling formula and descriptors

Shown earlier as an example (see Table 3-2), the final score scaling formula for this item is shown as a table, translating the range of possible final raw score totals to a four point scale with a final descriptor as shown in Table 5-5.

Table 5-5. Scenario 1, Phase 1: Final Score Scaling Formula

Raw score range	Final score	Final score descriptor
13–22	3	Strong evidence of recognising and understanding the central underlying analogy of the text.
8–12	2	Some evidence of recognising and understanding the central underlying analogy of the text.
1–7	1	Minimal evidence of recognising or understanding the central underlying analogy of the text.
0	0	No evidence of recognition or understanding the central underlying analogy of the text.

5.1.2.7 Score process, strategy, and design

The item responses for this scenario were scored by two raters. As the responses were expected to be relatively short and there were a larger number of items to score, the labour required for the scoring was minimised by recording only the scoring judgements themselves and not (for this scenario only) the specific sentence (or starting sentence) where the response specifically satisfied a given rubric criterion. While this prevents score reports for this short responses from associating specific scoring decisions with specific sentences, a five-sentence paragraph ensures that the

location of the information used to make the scoring decision is relatively clear and less critical to the evaluation of the scoring and to the feedback provided to students. In the remaining two scenarios, when scorers awarded a point, the point was stored in a record that established an association between the supporting rubric quality level definition and the sentence or sentence sequence that was the basis for the award.

Table 5-6. Scenario 1, Phase 1: Scoring Process

7. Scoring process	Strategy and design
(a) How scoring decisions recorded	For this relatively short-answer item, sentence–response associations are not required. For Scenarios 2 and 3, scoring decisions are recorded in association with a single sentence.
(b) How scoring data are used to produce a score report	The quality descriptors for the claim quality assigned, and the descriptors for any evidence points awarded, can be used to produce a report that shows these descriptions for each point awarded.
(c) how meaningful feedback is produced	The quality descriptors for each quality level or specific evidence point articulate why that factor is relevant to the overall score. All score reports can therefor include these quality level definitions as reflecting the rationale for the points awarded. Further, points not earned can be described as pathways to more complete responses. For claim points in Scenario 1, less than full credit can be explained by including the descriptor for the level of credit assigned by the scorer. Similarly, the total number of points of evidence reflect the degree to which the claim was supported with evidence in the text. Additional feedback will be available in Scenarios 2 and 3, where scoring for longer responses is supported by sentence-level sentence associations between score points from the rubric and the portion of the response responsible for the credit.
(d) Adjudication	In production use, scores by two scorers that differ by more than 1 point on the final scaled score are adjudicated by a third scorer who can review the detailed scoring judgements made by the initial scorers and construct a final score with its own individual subscore justifications for each point awarded.

5.1.2.8 Scoring process implementation

Scoring for all three scenarios of item responses and rubrics in this study was conducted by having two scorers working individually. Each scorer would begin by

reviewing the item and rubric details for a scenario, reviewing the scoring instructions that came with the spreadsheets (during the rubric development phase) and scoring web pages (in a scoring web application, for rubric testing phase). Each scorer would review an initial set of scored item responses prepared for instructional purposes, ask questions, and then undertake to score a test set for the first scenario and phase. After a review and discussion, scorers were then given additional guidance and more trial items to score or moved directly to scoring sets of items until their work was completed. All scorers scored at least one scenario including both development and testing phases, typically in batches of 40 item responses.

As one objective of this study was to examine IRR when using the RDF-based rubrics, there was no need to adjudicate differences in scores assigned by two raters when they were not equal. Instead, these differences were studied in detail to understand how the rubrics or feedback might be improved and to address the other research questions in the study. As the scoring process for all the scenarios was the same from an operational perspective, this aspect of the RDF for other items will refer to the description here.

5.1.2.9 Format and content of score reports

Rudimentary score reports created in Scenario 1 displayed each item response with associated meta-data (item and item response identification, rubric used) and points awarded by each scorer at each assigned quality level designation for each evaluative criterion or subscore. For Scenario 1, the scoring and feedback included a claim score for one of four quality levels (including zero) and an evidence score that reflected a tally of up to six points, one for any of six kinds of evidence. The reports showed for each of these summary-scored item responses,

- score points awarded for the claim (the claim subscore) and the associated quality level descriptor for that score;
- total evidence points recorded for any of the six evidence categories, and the total raw score and the final scaled score for each item; and
- the full text of the response for each item.

Ideally, end user reports developed for production use with this rubric would be based on the more robust scoring data capture used for Scenarios 2 and 3, which would then be able to associate specific feedback for specific sentences and kinds of evidence cited. As this was the first item scored, the automated system for capturing scoring data was not yet complete and most scores were captured on paper forms – so reports that could associate feedback with specific sentences could not be generated as sentence level scoring was not captured for Scenario 1.

While end user reports were outside the scope of this study, samples of the kinds of reports that could be generated from the information retained from the item delivery and scoring process are included in the discussion in the final chapter of this thesis.

5.1.2.10 Exemplars

For scorer training purposes, production use of this item would be supported by a robust set of scored examples at various levels of quality. An initial set of 40 scored items from this scenario was used in training for all scorers which provided a solid basis for understanding the scoring process and the structure of the kind of rubrics used in all three scenarios. The scoring for these examples was further refined when necessary for cases where the rubric was refined ahead of the test scoring phase of the project. As documented in the discussion of the development phase for each of the rubrics, these adjustments were relatively minor and generally reflected unexpected item response content or insufficient specificity in the quality level definitions themselves. Little changed as a training tool in the discipline of applying the rubric to a response.

Exemplars could be useful in production use as a way to recalibrate scorers whose judgements begin to drift from their original experience and consensus judgements. Some exemplars could be useful for student education, providing on a post-test basis additional samples of robust claim statements and explicit evidence citations in a now-familiar context. That said, sharing exemplars in many cases could compromise the utility of an assessment item for some future use cases (e.g., high-stakes assessment, or tests are designed to be time constrained, where pre-knowledge of an item could significantly advantage an examinee).

5.1.3 Holistic scoring results

Scenario 1 uses the scoring from the original holistic rubric (see 5.1.1) for this item for 40 selected item responses. These same items are to be scored using the newly devised RDF-based rubric as described in the next section. The original scores for the item responses included in Scenario 1, Phase 1, are presented below in a confusion matrix (Table 5-7) which shows how many item responses received each of the 16 possible combinations of scores from Rater 1 and Rater 2 (labelled H1 and H2).

Table 5-7. Scenario 1, Phase 1: Holistic H1 vs. H2 Score Comparison

Scenario 1 Phase 1 - h1 v h2 - holistic rubric -- 40 responses					
		criteria	total		
Agreement / Accuracy:		20	40	50%	
Adjacent Agreement:		40	40	100%	
Kappa with Quadratic Weighting			.95 Confidence Interval		
	Obs. Kappa	Std. Error	Lower Lmt	Upper Lmt	
	0.6537	0.1430	0.3735	0.9339	
H1/H2	**0**	**1**	**2**	**3**	
0	2				2
1		5	8		13
2		2	7	4	13
3			6	6	12
	2	7	21	10	**40**
0.8615	maximum possible quadratic-weighted kappa, given the observed marginal frequencies				
0.7588	observed as proportion of maximum possible				

Note. H1/H2 = human raters.

The holistic scoring for this item had a 50% interrater agreement, with no instances of scores being more than 1 point apart. The interrater agreement statistics for this initial set of item responses scored with the holistic rubric had a QWK of 0.6537. Adjacent agreement, defined as scores no more than 1 point apart, was 100%. This scoring serves as a baseline for comparison to the RDF scoring in later sections. The comparatively low QWK score that accompanies these high rates of agreement

reflects the small scale used (0 to 3) and significant possibility of chance agreement between two raters.

5.1.4 Initial RDF scoring results

The RDF scores for two raters scoring the Phase 1 set of 40 item responses for development purposes with the initial RDF rubric defined above (see 5.1.2) are shown in the confusion matrix (Table 5-8). The results were similar to the results for the holistic rubric, with some notable differences. The scorers using the holistic rubric show a 50% accuracy or exact agreement rate, as compared with a 63% exact agreement rate for the RDF scorers. With the RDF rubric, the most noticeable interrater disagreements were for the three item responses that showed a difference between the scorers of 2 points, a difference that did not occur using the holistic rubric which had a 100% "equal or adjacent" match rate. Another notable difference is the higher prevalence of zero scores from the RDF scoring. Both of these differences are analysed in the next chapter.

Table 5-8. Scenario 1 Phase 1: RDF C+E H1 vs. H2 Score Comparison

Scenario 1 Phase 1 - h1 v h2 - RDF rubric -- 40 responses					
		criteria	total		
Agreement / Accuracy:		25	40	63%	
Adjacent Agreement:		37	40	93%	
Kappa with Quadratic Weighting			.95 Confidence Interval		
	Obs. Kappa	Std. Error	Lower Lmt	Upper Lmt	
	0.7073	0.1302	0.4521	0.9625	
H1/H2	**0**	**1**	**2**	**3**	
0	6				6
1	3	8	3		14
2	3	3	6	2	14
3			1	5	6
	12	11	10	7	**40**
0.8049	maximum possible quadratic-weighted kappa, given the observed marginal frequencies				
0.8787	observed as proportion of maximum possible				

Note. C+E = claim + evidence; H1/H2 = human raters; RDF = rubric design framework.

5.1.5 Phase 1 Holistic versus RDF rubric results side by side

A summary description of the comparative results for the holistic and RDF-based rubric scoring for Scenario 1 is shown side by side in Table 5-9.

The holistic rubric produced a score distribution concentrated between the scores of 2 and 6, while the RDF rubric produced 4.5 times as many scores in the 0 to 1 range, as shown in Figure 5-1. As the focus of the holistic rubric was more on the form of the response and the command of English writing than on the content and substance of the claim. Accordingly, fewer zero scores were expected. That is, RDF scorers could award 0 scores for essays with reasonable writing if, for example, the response simply retold the story in the passage and failed to make a claim or cite evidence in response to the demands of the prompt, whereas the holistic rubric rewarded form without regard to specific content expectations.

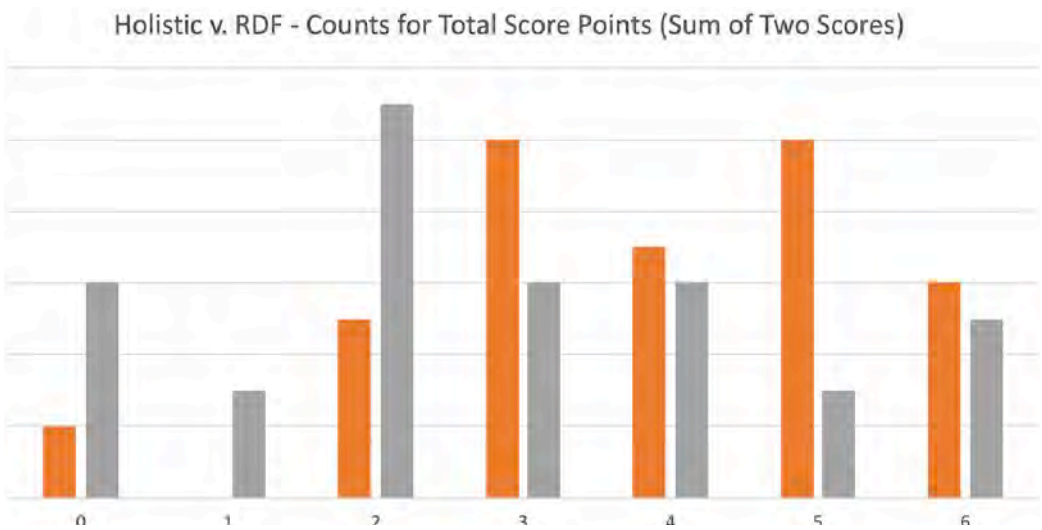
The score distribution across the 0- to 3-point scale used for both the holistic rubric and the final scaled score of the RDF rubric and the comparison of the IRR for the two rubrics are summarised in Table 5-9 and Figure 5-1.

Table 5-9. Scenario 1, Phase 1: Holistic vs. RDF Scoring Comparison

Interrater (H1 vs. H2) and distribution comparisons	Holistic rubric	RDF rubric
Number of item responses	40	40
Accuracy	50%	63%
Adjacent agreement	100%	93%
QWK	0.6537	0.7073
QWK standard error	0.1430	0.1302
Average score	1.93	1.4
Standard deviation	0.85	1.01
Median score	2	1

Note. H1/H2 = human raters; RDF = rubric design framework.

Figure 5-1. Scenario 1, Phase 1: Holistic vs. RDF Score Distribution



Note. These results are analysed in the context of the rubrics in Chapter 6. HOL = holistic; RDF = rubric design framework; WH = Winter Hibiscus.

5.1.6 Scoring analysis

The distribution of the scores from the RDF rubric scoring reflect, in their lower mean, and median, and the higher proportion of 0 scores, the more precise scoring criteria as compared to the holistic rubric criteria. The average number of sentences for responses in this group was 7.7; the median response length in number of sentences was seven, with three responses reaching 14 sentences and the longest response composed of 17 sentences. When scoring procedures were originally envisioned, the starting pool of responses had an average of only five sentences per response, so for this scenario, sentence-specific scoring was not undertaken.

IRR measures for the RDF scoring outcomes, as compared to the holistic scoring IRR, were both slightly above and slightly below the high level of exact agreement, adjacent agreement, and QWK as evidenced by the holistic scores; the RDF scores having 64% exact match, 93% adjacent agreement and a QWK of 0.7073 as compared with 50%, 100% and 0.6537, respectively. Overall, this level of IRR shows a rough equivalence between the two sets of scores.

Based on the feedback from scorers in the form of questions as to how to apply the rubric to specific responses, during scorer training, while scoring and from the post-scoring questionnaire (see Appendix I), and a review of differing scorers' evaluations of the same responses during the Phase 1 scoring exercise, some adjustments were made to the rubric that were designed to improve the quality and reliability of the scoring without compromising the potential for useful scoring feedback.

Three factors contributed to differences in scoring between the two scorers that related to the structure and content of the RDF rubric's specification, as described more generally in the prior section:

- ambiguity around the specified variations in less than complete recognition of the underlying analogy between Saeng and WH raised the issue of *partial claim complexity*;

- ambiguity around the degree to which evidence must be explicitly cited when a supporting observation is included in the response raised an issue around the *citation of evidence*; and

- ambiguity around dealing with extraneous or incorrect and contradictory content in a response that set itself off against evidence or claim aspects that were creditworthy raised issues with *extraneous content or misconceptions*.

Examples of rater feedback which was included on scorer notes that (contributed to these observations include:

- "Inventing analogy between winter hibiscus and Saeng's rough patch not supported by evidence." This reflected that a response claimed an analogy between Saeng and the WH, but not one which clearly fit into any of the development phase "partially correct" quality level definitions.

- "Mentioned 'start to become customed to her new country' - indicates adaptation, although not stated" is a case of the scorer trying to determine how much to infer in a response.

- "she wants to take the test again - this could indicate persistence but its linked it to the geese", one of many responses that contained analogies involving the migration of the geese that were unsuccessfully supported by evidence.

5.1.6.1 Partial claim complexity

In the case of partially correct statements of the underlying analogy of adaptation by Saeng and the WH, the claim subscore portion of the rubric was simplified to have only three values, not four: 16 points for recognition of the full analogy – that both WH and the immigrants need to adapt to a new place; 10 points for recognising a parallel between Saeng and WH of any sort, or for recognising the adaptation of Saeng *or* the WH; and 4 points for only citing the importance of determination or struggle. An example of partial claim complexity is illustrated in Figure 5-2.

Figure 5-2. Item Response 9523: Ambiguous Analogy

Item Response 9523:
In the concluding statement Saeng states that "when they come back In the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take the test again." The author concludes the story with this passage because she is using the hibiscus as an example of survival. The hibiscus is from another place like her, and she is using it to represent her well-being in the new place. In a way she is saying, If the hibiscus survives this winter, than so will I. I believe she uses this as a conclusion to depict her unsure nature towards her new home.

Figure 5-2 includes language that compares Saeng to the WH, noting they are both 'from another place'; it also explicitly connects the hibiscus survival to Saeng. The response has a focus around survival without any suggestion of adaptation, and mentions the geese returning and even Saeng's 'unsure nature towards her new home' without citing related evidence. In the initial trial scoring, one scorer rating this claim as worthy of 8 points; the other, 12. Among the 40 claim scores in this phase, there were also three cases of interrater difference on partial versus full credit scores for the claim, further suggesting a more general 'partial credit' for recognising comparison between Saeng and the WH as important, when not specific to the idea of adaptation, could be simplified and clarified for more consistent scoring and simplifying the associated feedback (e.g., 'recognition of an analogy or implied relationship between

Saeng and the WH without noting specifically the important parallel adaptation element').

5.1.6.2 Citation of evidence

To address confusion over citation of evidence, the language in the rubric was modified and rubric instruction was clarified to make explicit the requirement that evidence cited in the response should actually be used as intended - to support some aspect of the underlying analogy of the narrative - that Saeng or the WH had to evolve or adapt to survive; that this struggle would require work and sustained effort; or that change and adaptation was essential to survival because the environment had changed.

An example of ambiguous citation of evidence is illustrated in Figure 5-3.

Figure 5-3. Item Response 9343: Citation of Evidence

Item response 9341:
In the last paragraph of the story, the author concludes the story like that because the author means to say that the hibiscus is the symbol of overcoming obstacles. For example, Saeng's mother says 'Its flowers aren't pretty, but it's strong enough to make it through the cold months.' This is a symbol to overcome obstacles as a flower is strong enough to make it through the cold months, in which the weather is rough. In Saeng's case, she is a girl who will make it in a foreign country. Thus, in the last paragraph, when Saeng vows silently to herself that she will take the drivers test again when the snow melts (writers over) and the hibiscus is budding, she uses the hibiscus as a symbol of overcoming the obstacle of failing the test. That is why the author decided to end the story with that last paragraph.

This item response includes explicit parallel statements connecting Saeng and the WH, but the analogy is clouded by the focus on 'symbols' and the generic idea of 'overcoming obstacles' rather than any explicit suggestion of growth, change, or adaptation. The penultimate sentence is also nearly an exact quote from the passage, together with the second sentence; nearly half the content of this response (65 of 149 words) is text lifted directly from the passage. Within those copied aspects, retelling the story, one scorer saw no explicit analogy stated nor evidence to support it. Another scorer read the response to imply the relationship and saw three points of evidence in

the mentions of (a) WH is different from the one they knew – not as pretty; (b) WH strong enough to make it through the winter; and (c) Saeng's vow as evidence of determination. This and similar differences in scoring evidence suggested that the RDF rubrics should be as specific as possible about how to score responses in terms of implicit and explicit content. The degree to which a reader can or should make inferences from the response should be addressed to ensure a common understanding of what the scores mean, and should be consistently reflected in instructions, prompts, and support materials. At the same time, the overall weight of the score elements in this case—up to 16 points for the claim, and 6 points for evidence—serves to reinforce the CT focus of the exercise: If the claim is weak or missing key elements, it is lacking in a way that no amount of evidence can address.

5.1.6.3 Extraneous content and misconceptions

In the initial scoring, raters were sometimes confused about how to address material in a response that seemed to contradict the otherwise correct and valid (scorable) parts of the response. In other cases, misconceptions and errors in fact (e.g., misinterpretation of the item content) sparked questions from scorers about how such information should affect scoring (e.g., Should scores be lowered? Should points for evidence that is contradicted within a response be given less weight?).

An example of ambiguous citation of evidence is illustrated in Figure 5-4.

Figure 5-4. Item Response 9402: Misconceptions and Contradictions

Item Response 9402.
Including the bit about taking the test again when the 'hidiscus is budding' provides a sense of closeur and lesson to an otherwise open ending. Saeng fails the drivers test early in the story, then on her way home she buys a hibiscus. The reader really doesn't see a connection until this last paragraph. The hibiscus represents perserverence and strength. After a harsh winter this hibiscus can bloom again and and continue to live. Even after moving from Vietnam, and struggling to adapt, Saeng must recover, and keep trying. Just like the flower blooming again, so will Saeng. She will take the test a second time and pass. She will adapt to her environment, and keep going, just like the hibiscus.

This item response first asserts a claim about the reason for the closing paragraph, in response to the prompt. Subsequent text in the response shifts through alternative theories and different kinds of evidence, but ultimately crystalises the best solution in the final sentence.

Figure 5-5. Item Response 8870: Misconceptions and Contradictions

Item Response 8870.
I believe that the author concludes the story with this paragraph because the paragraph helps bring to story to a closing. In the paragraph she is basically saying when the geese return and my flower is budding I'll be prepared enough to take the test. She'll be prepared because she'll be adapted enough and she'll have experienced enough just like the plant and the geese. She Is comparing her life to the lives of the geese and the life of her flower.

Another item response that illustrates the potential for claims or evidence to include correct and incorrect elements simultaneously is shown in Figure 5-5. The response claims an analogy between Saeng and the geese (who migrate annually) as well as the WH (which has adapted to a new environment). Part of this response is spot on; the analogy with the geese is at best not obvious. The assertation that the geese and the WH gained experience or have adapted are not supported by evidence within the response, although the passage provides explicit support for the adaptation of the WH to Saeng's new home.

Different approaches to scoring such a response will be suitable for different purposes. The intent of the RDF rubric for this item is as much to educate as to evaluate, and so scoring instructions for this item with a formative focus were updated to explicitly communicate the preferred scoring of 'best interpretation possible' for explicit information, with reasonable discretion about implied aspects to a response. In other situations—for example, as part of a summative or mastery exam for CT—the RDF rubric should set expectations not only around what the response expectations are, but around what constitutes evidence of CT development and how contradictory or incorrect assertions, logic, and evidence should be addressed by the scoring.

5.1.7 Scenario 1 RDF rubric adjustments

The observations above contributed to some small changes to the first version of the rubric described earlier in 5.1.2. Partial claim complexity was addressed by a direct adjustment to the rubric between the development and testing phases, such that the middle tier of claim recognition was simplified, resulting in four, not five, score distinctions, as shown in Table 5-10 (as compared to the top Claim Subscore section of Table 5-4).

Table 5-10. Revised Claim Quality Level Definitions

3. Level descriptors	Quality level definitions (C+E)
Claim subscore:	Choose the score that best fits the claim in the response.
(a) Full credit/excellent claim: 16 points	(a) Full recognition of the underlying analogy
(b) Partial recognition of the underlying analogy: 10 points	(b) Recognition of the importance of adaptation for immigrants or WH; or a significant analogy between the two in their struggle, growth, or accommodation to their new surroundings
(c) Recognition of one aspect of adaptation or determination: 4 points	(c) Recognition of the importance of growth, struggle, determination, or adaptation for survival
(d) No recognition of the central underlying analogy	(d) No recognition of the central underlying analogy in the story or any of its major aspects

Note. C+E = claim + evidence; WH = Winter Hibiscus.

Citation of evidence issues were also addressed by clarifying the instructions included as part of the quality level definitions for the evidence subscore quality levels. This change is shown in a revised version of the evidence portion of Element 3 of the RDF in Table 5-11. The original version of the evidence sub-score "Quality level definitions" for this rubric are shown bottom Evidence Subscore in section of Table 5-4 in Section 5.1.2.

Table 5-11. Revised Evidence Quality Level Definitions.

3. Level descriptors	Quality level definitions (C+E)
Evidence subscore	One point for each. Evidence must be explicitly cited in support of reasoning or a claim statement, or directly and clearly implied.
(a) 1 point	Winter hibiscus is different from the hibiscus they knew before.
(b) 1 point	Winter hibiscus's flower not as pretty as the familiar one (different this specific way).
(c) 1 point	Winter hibiscus is strong enough, able to survive the winter/cold/snow (different this specific way from the familiar version).
(d) 1 point	Did what she had to do each day to give a good life to her child. Or Saeng is changing, beginning to experience her new environment as the new normal.
(e) 1 point	Persistence and determination are important. Adaptation is important, as survival is all important. Saeng is persisting/determined to survive.
(f) 1 point	The winter hibiscus here was still recognizable as related to the version they knew before; some aspects were the same: blood-red blossoms, five petals, long stamen, yellow pollen, feel of petal were 'exactly as expected'.

Note. C+E = claim + evidence.

Extraneous content and misconception issues were considered and discussed with all scorers. The decision for this exercise was to score for credit specified in the rubric. Had the rubric specified individual or general misconceptions for which negative credit would be applied, or even used according to a specific raw score calculation formula for negative credits awarded for specific misconception subscores, the rubric would have been applied. The rubric was clarified via instructions to the scorers that the task was to award credit for claims and evidence supporting those claims but to ignore for this exercise material that was incorrect, extraneous, or not relevant to the scoring. This direction was sufficient to resolve the issue in all cases, and no explicit change was made to the rubric as a result.

Scoring instructions were emphasised in personal interactions to reinforce the fact that credit should be awarded for proper evidence that was cited with explicit purpose or clearly implied purpose. Scorers were also warned that simple regurgitations of the story in the primary passage, with multiple quotes seemingly done as if in response to a request to summarise the material, should not be scored as if they were used to introduce evidence or further reasoning.

5.2 Scenario 2: Harriet Tubman: Claim and Evidence

In Scenario 2, HT item responses are scored with both the holistic rubric that was part of the original item materials and with a newly devised 'claim plus evidence' (C+E) rubric. The HT C+E rubric is drawn from the specific directive in the prompt to make a claim about the one quality of leadership that was most essential in enabling Harriet to guide the slaves to the North and the general directions to support the claim with evidence: 'and support your main ideas with evidence from both reading selections'. The writing intervention programme for which these assessments were designed explicitly ties AW to CT, as seen in the directions, instructions, and other item materials included as Appendix C.

5.2.1 Holistic rubric

The holistic rubric for the HT item is included as Appendix D. The rubric defines a 6-point scale with quality level descriptors as follows:

- 6, Exceptional achievement
- 5, Commendable achievement
- 4, Adequate achievement
- 3, Some evidence of achievement
- 2, Little evidence of achievement
- 1, Minimal evidence of achievement

At each of these quality levels, there are several quality definitions that address the expected content and quality with a variety of layered adjectives that in some cases characterise how the rater might evaluate qualities of the writing (*insightful/thoughtful* vs. *reasonably thoughtful* vs. *adequate* for a claim statement) and in other cases indicate the quantity of specific kinds of content coupled with related subjective quality descriptors (e.g., *discusses specific examples of several obstacles,* vs. *perceptively discusses examples of obstacles*, vs. *thoughtful discussion of examples*). Additional examples include *thoughtfully analyses a lesson*, *adequately analyses a lesson*, and *may provide a superficial lesson*; and *perceptively considers what characteristics, thoughtfully considers what characteristics*, and *considers what characteristics*. The complexity of the scoring challenge and the degree of variability and individual judgement that are required to assign a single score to a three- to five-paragraph response (averaging 19 sentences), is further illuminated by the

consideration of the number and range of these distinct quality definitions assigned at the various levels of the overall quality descriptor. Certainly, the assessment of each of the attributes captured by the individual quality level definitions will vary for the same response, and the rubric provides no guidance as to the relative weight of the varying dimensions. The number of quality definitions for each level of the unitary item response quality levels in this holistic rubric is shown in Table 5-12.

Table 5-12. Scenario 2: Number of Holistic Rubric Quality Definitions by Level

Quality level/descriptor	Number of quality definitions
6, Exceptional achievement	11
5, Commendable achievement	11
4, Adequate achievement	10
3, Some evidence of achievement	10
2, Little evidence of achievement	8
1, Minimal evidence of achievement	7

In the actual scoring of the item responses with the holistic rubric (as will be shown in Section 5.2.3, Table 5-20), the two raters' scores were in exact agreement or were only off by only 1 point for all 40 items scored on the six-point scale. While this is potentially surprising, I noted that the scoring judgements provided by the staff that run the writing intervention programme could be influenced by a number of factors, including

- natural collinearity of the observed traits in the item responses themselves;
- shared practice and training; or
- the use of exemplars and/or shared values among the raters.

Any of these factors could tend to equalise the judgements required to assign a single overall score when multiple levels of score descriptors are satisfied by different parts of a single response.

5.2.2 RDF rubric

The RDF-based rubric devised for the HT item in Scenario 2 is shown below, addressing the 10 aspects of the rubric design framework set forth in Section 3.4.

5.2.2.1 High-level rubric definition

The RDF-based rubric for this scenario is a two-part, C+E rubric similar to the RDF rubric for Scenario 1. The prompt for this item requires the student to make a claim and support it with evidence. Accordingly, this scenario has two subscores, one for claim and one for evidence, as in Scenario 1. Based on the instructions provided, which called for the choice of a 'most important leadership trait', the directive for this challenge was to support the claim with evidence from the passages and to illustrate the claim with multiple examples. As a result, the balance or relative weights of the subscores for this rubric reflect the requirements of the prompt, with a distribution between the points awarded for the claim and the points awarded for evidence shifted to reflect the greater emphasis on evidence. In particular, the challenge may be supported by the student's explanation of why a particular trait was chosen and further illustrated by citing direct support from several examples in the text. This rubric assigns an overall CT score based on the combination of the quality of the claim and the use of evidence, yielding a score on a scale of 0 to 6, comparable to the holistic score range.

For this detailed RDF rubric, the initial choice was to allocate 4 points for a clear and comprehensive claim, leaving room for partial credit for claims that had flaws (as explained more fully below in the quality level definitions). This rubric allocates a maximum of 8 points for evidence, giving twice as much weight to evidence as to the claim. This reflects that the rubric accepts any of the seven traits defined in the auxiliary passage as an answer, placing more focus on the form and completeness of the claim (e.g., does it specify one single most important trait, as directed, or merely cite an important trait?) while requiring several examples to support whatever trait is chosen. These evidence claims could associate the reason for the choice with many possible events and circumstances in the primary passage; allowing up to eight separate bits of evidence was a useful way to get a comparative measure of levels of evidence across responses of very different levels of depth and engagement. The

initial plan was to report these scores on scale comparable to the original 6-point scale.

Table 5-13. Scenario 2 Phase 1: High-Level Rubric Definition

1. Rubric definition	HT RDF critical thinking (C+E)
(a) Construct: skill, knowledge or capability measured	Critical thinking: Understanding a historical account described in story form and an informational text (on leadership), and synthesise these two so as to draw inferences from which to make a claim and support it with evidence.
(b) Audience	Student with 8 to 10 years of schooling who can read English as expected for this grade level.
(c) How assessed	The item consists of (a) a story passage illustrating the work of Harriet Tubman to help free slaves from the South during the time of the American Civil War; (b) informational text on leadership traits; (c) a challenge prompt that requires understanding and synthesising information from both sources; and a set of instructions for addressing the challenge in the desired response. The challenge requires the student to make a claim and to support the claim with evidence.
(d) How scored	A proper claim that addresses the demands of the prompt will make a claim based on the definitions in the auxiliary passage and support the claim with information and examples from the primary passage. A well-formed claim will be awarded 1/3 of the total potential score for the item, and citation of evidence will account for the other 2/3 of the scores. The prompt includes suggestions for the kinds of evidence expected that will focus the scoring work.
(e) Security/disclosure	The selection of a specific trait or specific evidence for a given example will not by itself compromise the utility of this item, but if the goals for its use include some sort of comparative ranking or performance among a population of students, students with pre-exposure to the challenge or exemplary responses will have an advantage over students without such foreknowledge—particularly in settings that are time limited. For formative use, these concerns are muted beyond the usual concerns that some students will memorise rather than learn and internalise the knowledge.

1. Rubric definition	HT RDF critical thinking (C+E)
(f) Anticipated use	An exercise to gauge a student's ability to synthesise data from multiple sources, make inferences, articulate claims and cite evidence from an intermediate-level text, both historical and informational. With robust feedback, students can learn from score reports how to address any specific deficiencies identified in their response.

Note. C+E = claim + evidence; HT = Harriet Tubman; RDF = rubric design framework.

5.2.2.2 High-level item structure

The item used in Scenario 2 is composed of a primary passage that contains a narrative describing a historical figure in action; an auxiliary passage that is an informational text on leadership traits; a prompt that requires an examinee to understand and synthesise information from both artefacts; and instructional materials that provide guidance on how to organise and present appropriate content to respond to the demands of the prompt.

Table 5-14. Scenario 2 Phase 1: High-Level Item Structure

2. Item definition	HT item: Passages, prompt, and instructions
(a) Primary passage	The item uses a passage, 'The Railroad Runs to Canada', from *Harriet Tubman: Conductor on the Underground Railroad,* by Ann Petry.
(b) Auxiliary passage	The item uses a short paper, 'Seven Qualities of a Good Leader', by Barbara White.
(c) Prompt	The item includes a one-page prompt with background, writing instructions, a prompt question, and instructions for both the body and the conclusion of the desired response.
(d) Instructions	The instructions come in the form of a four-page document with heading 'Pathway Project Reading and Writing Assessment'.

Note. Item is included in Appendix C. HT = Harriet Tubman.

5.2.2.3 Scoring criteria and level definitions

The RDF criteria require explicit scoring (or evaluative) criteria, level descriptors and level definitions. The initial RDF for Scenario 2 has the following scoring criteria, level descriptors, and definitions. The primary scoring factors are reflected in the

points awarded for synthesising the information provided into a clear and direct claim in response to the prompt. Points for partially formulated claims are allowed. A fully conforming claim statement identifies a single trait from the several defined in the auxiliary passage as the most important to HT's success. The quality level descriptions and definitions for the claim subscore are shown in Table 5-16. Evidence supporting this claim is also required, and based on the instructions, specific aspects of the item content can be anticipated as contributing to the supporting evidence. The evidence level descriptions and quality level definitions are shown in Table 5-17. Claim and evidence scores are also calculated and combined for a final score on a 1- to 6-point scale, comparable to the score report from the holistic scoring.

Table 5-15. Scenario 2 Phase 1: Scoring (Evaluative) Criteria

3. Scoring criteria	HT Item: RDF C+E Rubric
(a) Claim subscore	Ability to synthesise information and articulate a claim
(b) Evidence subscore	Ability to articulate supporting evidence for reasoning

Note. C+E = claim + evidence; HT = Harriet Tubman; RDF = rubric design framework.

Table 5-16. Scenario 2, Phase 1: Claim Subscore Quality Level Definitions

3. Level descriptors	Quality level definitions (C+ E)
Claim subscore	
(a) Full credit: 4 points	(a) A single trait from the auxiliary passage is identified as the most important leadership trait that helped HT succeed.
(b) No credit: 0 points	(b) The response fails to identify the most important leadership trait that helped HT succeed.
(c) Most credit: 3 points	(c) A single trait is identified and described as important to HT's success, but an explicit 'single most important' declaration is absent.
(d) Partial Credit 1 or 2 points	(d) Responses that identify multiple traits, or that identify a trait or traits not part of those identified in the auxiliary passage, may get partial credit depending on the clarity of their claim and the coherence of their reasoning.

Note. C+E = claim + evidence; HT = Harriet Tubman.

Table 5-17. Scenario 2, Phase 1: Evidence Quality Level Definitions

3. Level descriptors	Quality level definitions (C+ E)
Evidence subscore 1 to 2 points	Evidence to support the claim for a most important trait Most examples will be 1 point; strong or well-developed citations of evidence and reasoning can be 2 points.
Evidence such as HT's	Actions or choices that manifest, support or reflect the indicated trait as asserted in the response Observations of differences between HT and her followers that contribute to the quality/leadership success Observations of similarities between HT and her followers, common characteristics that (as explained by the writing) contribute to the quality/leadership success

Note. C+E = claim + evidence; HT = Harriet Tubman.

5.2.2.4 Subscale score calculation formula

The claim score is marked on a 0 to 4 scale based on the factors above. Evidence used in supporting the claim generates 1 or 2 points per citation, to a maximum of 8 points. The claim score and the sum of the evidence points represents subscore totals; added together they represent a total scaled score.

5.2.2.5 Final raw score formula

The raw score formula sums the subscores, resulting in scores between 0 and 12.

5.2.2.6 Score scaling formula and descriptors

The final score scaling formula divides the 0 to 12-point scale score by 2, rounding up to a whole number and using the whole number result, giving a final scaled score of 0 to 6. The scores then rest on the same scale as the original 6- point scale and can use the same six quality level descriptors as shown in Table 5-18 below - recognising that the value from scoring in this case is derived from the specific identification of response elements that are credited with claim or evidence qualities that respond to the prompt, rather than a holistic assessment of a rater across a dozen or more factors of unspecified weight.

Table 5-18. Scenario 2, Phase 1: Final Score Descriptor

Final score	Final score descriptor
6	Exceptional achievement
5	Commendable achievement
4	Adequate achievement
3	Some evidence of achievement
2	Little evidence of achievement
1	Minimal evidence of achievement
0	Nonresponsive

5.2.2.7 Score process, strategy, and design

The item responses for Scenario 2 are to be scored by two raters, who assign score points for specific evaluative criteria (subscores) to specific sentences in the response text. The purpose of recognising a location for each score point is to associate the feedback for a given point with a specific part of the response to enhance the utility and meaning to the feedback. When the exact locus of an observation, claim, or bit of reasoning is developed over many sentences or in a clause within a single sentence, the entire first sentence that contributes to the rater's assessment that something scorable was found is chosen as the location for the purpose of pointing the score report user to the proximate source of the contribution.

Table 5-19. Scenario 2, Phase 1: Scoring Process

7. Scoring process	Strategy and design
(a) How scoring decisions are recorded	Scoring decisions for either claim or evidence points were recorded in association with a specific sentence in the response. These indicate the number of points awarded (1 or 2 for evidence points; 0 to 4).
(b) How scoring data are used to produce a score report	With the point-to-sentence associations captured during scoring, scoring reports can be generated that enumerate the value and location of each claim or evidence point awarded to a response.
(c) How meaningful feedback is produced	By having distinct subscore points awarded to specific elements of the response, and recording these associations, the rationale for the scorers' judgements can be reported and analysed. Students will see where credit was awarded and aspects of the rubric that were satisfied (and not), providing explicit and implicit feedback to each aspect of the response.
(d) Adjudication	Third scorer adjudication will be used for discrepant (more than 1 score point on a 6-point scale) scores.

5.2.2.8 Scoring process implementation

The item scoring process used across the three scenarios was described in Scenario 1 (Section 5.1.2.8). Scoring for Scenarios 2 and 3 differed only by the capture at the point of scoring the specific sentence within an item response where the text satisfies a specific rubric criterion (or quality level definition). If the relevant text spans multiple sentences, the association is with the first sentence in the group. Such associations are stored for every point awarded by the scorer. With this additional information captured, item responses scoring can be more discretely examined, and score reports can provide more information to students, teachers, and other audiences as described in the following paragraphs.

5.2.2.9 Format and content of score reports

The score reports created for Phases 2 and 3 add a significant layer of detail to the scoring by showing specifically which part of a response satisfied which part of the rubric, rendering scoring more explainable and providing explicit feedback to students. Unearned points can help explain deficiencies in responses, while specifying which sentence met which quality level criterion affords students a learning

opportunity and gives instructors more accessible information about what a student knows and can do. Although end user reports that include diagnostic and feedback detail were not developed as part of this study, detailed score reports were generated to detail all item response content to rubric connections. With these detailed reports, differences between scorer decisions on the same item responses were revealed, which allowed the rapid evolution of rubric improvements to address situations where either a rubric quality level definition or an item response itself revealed ambiguity that refinements to the rubric could clarify. Although end user reports were outside the scope of this study, examples of the sort of reports enabled by capturing this scoring data, with or without sentence-level scoring detail, are included in the final chapter of this thesis.

5.2.2.10 Exemplars

As explained in Scenario 1 (Section 5.1.2.10), exemplars were not created and formalised during this rubric development study but would be an excellent addition to the item and rubric information used for items of this kind planned for production use. Items selected for training in this study, which included sample item responses across the range of possible scores, would provide a useful starting point for establishing a set of exemplars for this item prior to large-scale deployment.

5.2.3 Holistic scoring results

Scenario 2 uses scoring from the original holistic rubric (see 5.1.3) for this item for 40 selected item responses. These same items will also be scored using the newly devised RDF-based rubric as described in the next section. The original scores for the item responses in Scenario 2, Phase 1, are presented below in a confusion matrix (Table 5-20), which shows how many items received each of the many possible combinations of scores from Rater 1 and Rater 2 (labelled H1 and H2).

Table 5-20. Scenario 2, Phase 1: Holistic H1 vs. H2 Comparison

HT Item - h1 v h2 - Original Holistic Rubric -- Initial 40 Item Responses								
		criteria	total					
Agreement / Accuracy:		22	40	55%				
Adjacent Agreement:		40	40	100%				
Kappa with Quadratic Weighting			.95 Confidence Interval					
Observed Kappa		Std. Error	Lower Lmt | Upper Lmt					
0.9080		n/a	n/a | n/a					
H1/H2	**0**	**1**	**2**	**3**	**4**	**5**	**6**	
0								
1		2	2					4
2		2	4	1				7
3			3	4	2			9
4				2	3			5
5					2	6	2	10
6						2	3	5
		4	9	7	7	8	5	**40**
0.9796	maximum possible quadratic-weighted kappa, given the observed marginal frequencies							
0.9269	observed as proportion of maximum possible							

Note. H1/H2 = human raters; HT = Harriet Tubman.

The writing intervention programme of which this assessment is a part is itself part of a long-running programme at many universities across the United States. This writing programme is run annually and is staffed by a dedicated group of writing and English instructors. Their close agreements on scoring, while perhaps surprising given the breadth and range of the written rubric, reflects years of practice in the application of their materials to the course.

The holistic scoring for these item responses had a 55% interrater agreement, with no instances of scores being more than 1 point apart across the entire 6-point scale. Accordingly, the interrater agreement statistics for this initial set of item responses scored with the holistic rubric had a QWK of 0.9080. Adjacent agreement, defined as scores no more than 1 point apart, was 100%. This holistic scoring serves as a baseline for comparison to the RDF scoring in later sections.

5.2.4 Initial RDF scoring results

The RDF scores for two raters scoring the same 40 item responses as were scored with the holistic rubric in the prior section were also scored with the initial RDF rubric defined above (5.2.2). The results of this scoring are shown in the confusion matrix (Table 5-21) below. The RDF scoring for HT items with the claim plus evidence rubric were broadly similar (but less in agreement) to the scoring with the holistic rubric, with lower levels of exact and adjacent agreement (40% and 78%, as compared to 55% and 100%). The RDF scorer comparison showed a reasonably high level of overall agreement (QWK of 0.7919, though lower than the very high QWK observed for scorer agreement of 0.9269 for the holistic rubric). The RDF IRR measure was somewhat lower than for the holistic score but still above thresholds typically used to validate operational items for production use in high stakes testing (e.g., a QWK exceeding 0.70, per Williamson, Xi and Breyer, 2012, p. 7). In comparing the scores using the two rubrics across their range of results, a general drift toward lower overall scores from the RDF rubric is visible, expected by virtue of the more exacting scoring criteria as compared to the holistic rubric. The scoring results are more fully reviewed and analysed in the following sections.

Table 5-21. Scenario 2, Phase 1: RDF Scorer Comparison

Scenario 2 - h1 v h2 - Initial RDF Rubric -- 40 Item Responses								
			criteria	total				
Agreement / Accuracy:			16	40	40%			
Adjacent Agreement:			31	40	78%			
Kappa with Quadratic Weighting				.95 Confidence Interval				
Observed Kappa		Std. Error		Lower Lmt | Upper Lmt				
0.7919		n/a		n/a | n/a				
H1/H2	**0**	**1**	**2**	**3**	**4**	**5**	**6**	
0	7	2		1				10
1		2		1				3
2		1	1	3		1		6
3			1	1			2	4
4				3	1	3	2	9
5			1	1			2	4
6							4	4
	7	5	3	10	1	4	10	**40**
0.8905	maximum possible quadratic-weighted kappa, given the observed marginal frequencies							
0.8893	observed as proportion of maximum possible							

Note. H1/H2 = human raters; RDF = rubric design framework.

5.2.5 Phase 1 holistic Versus RDF rubric results side by side

Table 5-21 reveals three clusters of two score pairs with difference of 2 or more points (score pairs outside the exact and adjacent central band) for further exploration. A summary description of the comparative results for the holistic and RDF-based rubric scoring for this first phase of Scenario 2 is shown in Table 5-22. These results are analysed in the context of the two rubrics in the following sections.

As shown by the score distributions in Figure 5-6, the RDF rubric resulted in somewhat lower scores for this sample of 40 items (e.g., the mean and median were both significantly lower). This was expected insofar as the RDF rubric is focused on a claim and evidence that satisfies specific criteria, whereas the holistic rubric considers a broader range of concerns and includes such factors as general writing mechanics. This difference is highlighted in Figure 5-6, which shows that scores 3 and 5 (on a 6-point scale) are the most frequently assigned scores by holistic rubric graders, whereas

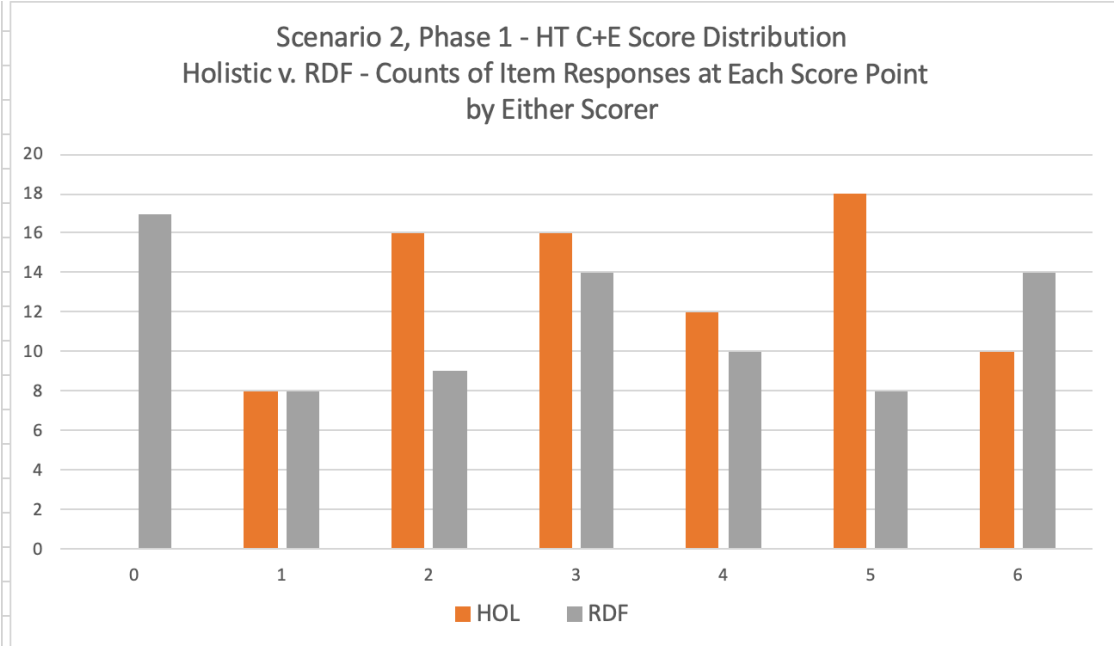
the score most frequently assigned on these same responses by RDF rubric scorers was 0.

Table 5-22. Scenario 2, Phase 1: Holistic vs. RDF Scoring Comparison

Interrater (H1 vs. H2) and distribution comparison	Holistic rubric	RDF rubric
Number of item responses	40	40
Accuracy	55%	40%
Adjacent agreement	100%	78%
QWK	0.9080	0.7919
QWK standard error	n/a	n/a
Average score	3.58	2.9
Standard deviation	1.57	2.10
Median score	3.5	3

Note. H1/H2 = human raters; QWK = quadratic weighted kappa; RDF = rubric design framework.

Figure 5-6. Scenario 2, Phase 1: Holistic vs. RDF Score Distribution Chart



Note. C+E = claim + evidence; HOL = holistic; HT = Harriet Tubman; RDF = rubric design framework.

5.2.6 Scoring analysis

As with Scenario 1, the distribution of the RDF rubric scores for Scenario 2 reflects, in their lower average and median score values, the results of applying more specific and rigorous scoring criteria to the responses. In addition, and again as was seen in

Scenario 1, the RDF scoring results in a greater number of 0 or very low scores reflect responses that seem to ignore the demands of the prompt, typically in the form of story-telling responses that essentially summarise the passages rather than respond to the prompt or challenge question.

A detailed review and comparison of subscores (points awarded for claim and evidence) focused on item responses for which subscore differences were greatest. The 10 item responses with more than a 1-point subscore difference in either category included nine responses with a 2 or more point difference in evidence subscores and one item response with such differences in both claim and evidence subscores.

The differences in evidence score generally had their roots in poor writing, making scoring judgements more difficult, as some scorers were willing to infer more than others. Three specific kinds of challenges were noted during the review. The first and most frequent was the question of how explicitly evidence needed to be called out as evidence, rather than simply stated as support, to earn credit. The second challenge in scoring evidence related to the role of reasoning and when this should be part of the credit for the evidence citation. Finally, the one instance of a significant difference in claim scores assigned went directly to the issue of 'if asked for a single trait, is it acceptable to respond with three traits, or in some other way', which different scorers might resolve differently in the absence of specific directions.

The three examples below illustrate these scoring challenges that inspired minor adjustments to the specificity of the rubric:

- How strictly the response should adhere to the specific demands of the prompt, particularly with regard to the claim (Item Response 3572);

- Ambiguity around how explicitly evidence must be cited when used to support the claim (Item Response 3661); and

- How explicitly evidence must be connected by reasoning to a claim, particularly when errors in language use, spelling, and grammar obscure the point being made (Item Response 3536).

An example noted by scorers in their commentary on scoring for these responses is that two scorers said, with slightly different wording, “partial credit for claim of courage” –when the rubric did not specifically address the selection of a trait other than the seven defined in the auxiliary passage. One gave 1 out of 4 points for the claim; another gave 3. As a result, the claim score quality levels were augmented to provide common guidance for this and related situations.

5.2.6.1 Demands of the prompt

This challenge is illustrated by Item Response 3572 in Figure 5-7, in which, rather than failing to identify a specific trait, or to identify more than one trait, the response invented an entirely new trait—a situation not anticipated explicitly in the rubric.

Figure 5-7. Item Response 3572, Demands of the Prompt

Item Response 3572.

Harriet Tubman was a brave woman conducting the underground railroad. From the southern plantations she guided slaves into freedom. There were big concequences if she got caught, but she still did it anyway.
I think the most essential quality of leadership that enabled Harriet to guide slaves into freedom was curage. It takes alot of curage to do what she in her time. Most people today would not have the curage to do what Harriet Tubman has done for the slaves.

The instructions and the prompt itself (as reproduced in Appendix C) indicated, in all caps and bold, that the student was to read the Barbara White article and make a claim about one quality of leadership. The rubric’s Claim Subscore Quality Level Definitions in Table 5-16, which specifically addressed the selection of two or more traits, was accordingly augmented to specifically note that selecting traits not part of the Barbara White essay was nonresponsive.

5.2.6.2 Implicit use of evidence

This challenge is illustrated by the scoring for Item Response 3661, the second paragraph of which is illustrated in Figure 5-8.

Figure 5-8. Item Response 3661, Implicit Citations

Item Response 3661, paragraph 2:

Harriet tubman best quality of leadership is steadfastness. We can learn from her to always keep a cool head and to keep on going when there is a bad situation. Harriet and her followers share common they were all slaves, they all went to get out of danger. They came throught obstacles like when they had to walk to get some shelter and food. The first farm they came didn't let them get in ad they had to walk tired exhausted and hungry more so they could make it to the other.

This paragraph follows a long introductory paragraph that includes a sentence that begins 'The most quality of leadership was most essential in enabling Harriet to guide the slaves to the north was steadfastness because …'. This follow-on paragraph shown in Figure 5-8 begins with a faint echo of that claim, albeit one that does not parse well. The sentences that follow are somewhat broken but could be read as conveying by example why HT had to be steadfast. It was not marked off with a specific citation such as 'An example of when HT had to be steadfast…', but the proximity and cadence of this second paragraph can reasonably be seen as citing evidence to support the claim. Equally, it could be seen as a somewhat jumbled bunch of nonsense whose exact meaning and intent could only be guessed at. The two evidence scores for this item reflected more or less these two extremes.

Many of the writing samples that gave rise to the most challenging scoring issues demonstrated a limited command of the standards and conventions of standard written English. After more consideration, and given that the focus of the assessment is to have the student demonstrate their ability to make a claim and support it with evidence, I decided to improve the language of the evidence quality level definition (see Evidence Quality Level Definitions in Table 5-17) with regard to the evidence subscores. The desired result is to make specific that if the writing is not clear or evidence is not cited explicitly or implicitly by the specific context, then the scorer should not make assumptions about the writer's intent or meaning but score what is directly on the page—which may well mean that specific evidence that could be cited for credit will not be credited if it is not cited.

5.2.6.3 Reasoning and evidence

This challenge is illustrated by the scoring for Item Response 3536, produced in part below. The final sentence of the paragraph provides the reasoning that illuminates the use of the observation that HT could safely sleep with her followers, who she at times had to threaten, to support the argument that confidence was essential to trust and, in turn, to her effectiveness as a leader.

Figure 5-9. Item Response 3536, Reasoning and Evidence

Item Response 3536, from second paragraph (of four).

Also, they had trust with each other. Like said in the biography, During the trip, Harriet would suddenly hall asleep. The slaves could have run away, go back, leave her, killed her, or stolen something, but they didn't; they waited patiently until she awaken. It was the trust that build up from the confidence that maintained them together.

In many instances reasoning statements were counted as evidence by some scorers and not others. As a result, the evidence quality level definitions in the rubric (Table 5-17) were enhanced to make explicit that such reasoning statements could be germane and counted toward evidence in a response.

5.2.7 Scenario 2 RDF rubric adjustments

As noted in the examples above, the three kinds of errors noted in the analysis result in additional notations in the RDF quality level definitions for both claims and evidence to improve overall consistency in score and support the assessment's focus on CT-related construct elements. To recap, they were additional annotations on the claim quality level definitions as described and shown in Table 5-23 (which updates Table 5-16), and additional annotations on the evidence quality level definitions as described and shown in Table 5-24 (which updates Table 5-17).

The nature of the rubric adjustements described above include:

Claim subscore changes:

- Claims for traits *not* included in the secondary passages are not responsive to the demands of the question, and so not credited.

Evidence subscore changes:

- While citations of evidence are not required to include language such as 'the evidence of X come from the observation Y because of Z', evidence should be clearly connected to its use and reasoning and not depend on the readers' own intuition.

- Reasoning itself can be cited as part of the evidentiary support to the claim.

These changes are reflected in the 'additional notes' section at the bottom of the updated RDF element tables below.

Table 5-23. Scenario 2, Updated Claim Quality Level Definitions

3. Level descriptors	Quality level definitions (C+E)
Claim subscore	
(a) Full credit: 4 points	(a) A single trait from the auxiliary passage is identified as the most important leadership trait that helped HT succeed.
(b) No credit: 0 points	(b) The response fails to identify the most important leadership trait that helped HT succeed.
(c) Most credit: 3 points	(c) A single trait is identified and described as important to HT's success, but an explicit 'single most important' declaration is absent.
(d) Partial credit: 1 or 2 points	(d) Responses that identify multiple traits, or that identify a trait or traits not part of those identified in the auxiliary passage, may get partial credit depending on the clarity of their claim and the coherence of their reasoning.
Additional notes (added for Phase 2)	Claims for traits not included in the secondary passages are not responsive to the demands of the question.

Note. C+E = claim + evidence; HT = Harriet Tubman.

Table 5-24. Scenario 2: Updated Evidence Quality Level Definitions

3. Level descriptors	Quality level definitions (C+E)
Evidence subscore: 1 to 2 points	Evidence to support the claim for a most important trait Most examples will be 1 point; strong or well-developed citations of evidence and reasoning can be 2 points
Evidence such as HT's	Actions or choices that manifest, support or reflect the indicated trait as asserted in the response Observations of differences between HT and her followers that contribute to the quality / leadership success Observations of similarities between HT and her followers, common characteristics that (as explained by the writing) contribute to the quality/leadership success
Additional notes (added for Phase 2)	While citations of evidence are not required to include language such as 'the evidence of X comes from the observation Y because of Z', evidence should be clearly connected to its use and reasoning, and not depend on the reader's own intuition. Reasoning itself can be cited as part of the evidentiary support to the claim if it does so in the judgement of the scorer.

Note. C+E = claim + evidence; HT = Harriet Tubman.

5.3 Scenario 3: Harriet Tubman: Narrative Elements

In Scenario 3, HT item responses are compared when scored with both the holistic rubric that was part of the original item materials and with a newly devised 'narrative elements' rubric. The HT narrative elements rubric is drawn from the directives within the item training, instructions, and prompt that defined specific content requirements (or suggestions) for the response: (a) identify the key quality of leadership; (b) discuss why it was so important to her and the other slaves' survival; (c) discuss how HT's response to life-threatening situations was similar to the reactions of her followers; (d) discuss how HT's response to life-threatening situations was different from the reactions of her followers; (e) what HT has in common with her followers; (f) what differences allow HT to emerge as a leader; and (g) what general lesson can be taken from this story. The idea that these elements were essential parts of the task was initially defined by the first four pages of instruction used in classroom exercises of

which this assessment is a part (see 'Pathway Project Reading and Writing Assessment', Appendix C), reinforced by the classroom-based exercises where each of these narrative elements were discussed and explored (and part of written assignments), and finally restated in the section of the prompt entitled 'In the body of your essay' immediately above the text box used for the writing prompt.

By including Scenario 3 in this study, with its more generalised narrative elements rubric requirements, I consider the relative success in applying the RDF framework to content-centric, item-specific content beyond the common core of CT attributes related to claim and evidence. This provides an additional lens through which to review the quality of an AW assignment, the impact of specific aspects of the instruction, and more information about the effectiveness of the writing intervention as a whole. A consideration of the similarities and differences between scores from the distinct rating exercises is therefore included among the concluding analyses conducted and discussed in later chapters.

As there are seven distinct narrative elements called out for inclusion in this writing exercise, this rubric is also referred to as the Narrative Elements A–G Rubric, or A–G for brevity.

5.3.1 Holistic rubric

The holistic rubric for the HT item, also used in Scenario 2, is included as Appendix D. This rubric was reviewed in detail in 5.2.1, for Scenario 2, and also serves as the baseline against which RDF scoring for these narrative elements will be compared.

5.3.2 RDF rubric

The RDF-based rubric devised for the HT item in Scenario 3 is shown below, addressing the 10 aspects of the rubric design framework set forth in Section 3.4.

5.3.2.1 High-level rubric definition

The RDF-based rubric for this scenario is a seven-part 'narrative elements' rubric designed to assess the degree to which the item responses follow the guidance and instruction for the content. These instructions accompany the AW coursework and are included in the class instruction leading up to the item administration. They were also

repeated within the body of the prompt itself. They are meant to encourage the development of a range of ideas to be synthesised and incorporated into the response. This rubric checks for this breadth of topic coverage, awarding progressively more points to responses that include proportionally more aspects of all seven content categories. Some topics—such as how HT is different in ways that contribute to her success—warrant more focus than the guidance directed at encouraging the response to include lessons learned as they more directly contribute to the AW subject itself. The original rubric was scored on a 1 to 6 scale. This rubric awards 1 or 2 points in up to seven categories, for a maximum of 12 points, with a final scaling operation to place the score on a 6-point scale similar to the holistic score scale. This allows some level of comparability to both the RDF C+E rubric and to the original holistic, more general quality-of-writing rubric.

For this detailed RDF rubric, the initial choice was to allocate 1 or 2 points for each of seven categories, with two of the categories least directly contributing to the choice of trait and differentiating factors—the 'lessons learned' and 'how is HT most like her followers'— limited to 1 point. The other factors were weighted at 2 points to create a maximum of 12 points.

Table 5-25. Scenario 3, Phase 1: High-Level Rubric Definition

1. Rubric definition	HT RDF narrative elements (A–G)
(a) Construct: skill, knowledge or capability measured	The rubric assesses the degree to which the student included the full range of recommended content types in the response provided. This can be used to assess the correlation between stronger responses and following these specific recommendations and also provide some feedback to the teachers as to the impact of their instructional materials on the students' decisions about their responses.
(b) Audience	Students with 8–10 years of schooling who can read English as expected for this grade level.

1. Rubric definition	HT RDF narrative elements (A–G)
(c) How assessed	The item consists of a story passage illustrating the work of Harriet Tubman to help free slaves from the South during the time of the American Civil War; informational text on leadership traits; a challenge prompt that requires understanding and synthesising information from both sources; and a set of instructions for addressing the challenge in the desired response. The instructions define seven kinds of content and ask the student to consider using the full range of content types in their responses.
(d) How scored	Scorers assign points to content by identifying which of the seven content types are present in the response and score a 1 or 2 based on the number and depth of observations of each type. These point allocations are attributed to individual sentences. For the expression of an idea that is composed of contributions from multiple sentences, the points will be associated with the first sentence of the group.
(e) Security/disclosure	The rubric that generally reveals that the response will be scored in part based on the degree of coverage provided in the response of the recommended topics in the content of the response should not impact the utility of the assessment. Detailed disclosure of the specific weights of different categories or how the narrative elements are defined, beyond what is already provided in the training and item materials, is not necessarily harmful, nor likely useful.
(f) Anticipated use	This item is both an assessment of information synthesis and related critical thinking activity, as well as a mechanism to assess the impact of a particular set of instructional interventions in support of the argumentative writing task itself. This exercise overall will help gauge a student's ability to synthesise data from multiple sources, make inferences, articulate claims, and cite evidence from an intermediate-level text, both historical and informational.

Note. HT = Harriet Tubman; RDF = rubric design framework.

5.3.2.2 High-level item structure

The high-level item structure is the same as described in Section 5.2.2.2; the table of information containing this description is repeated here for convenience (Table 5-26).

Table 5-26. Scenario 3, Phase 1: High-Level Item Structure

2. Item definition	HT Item: Passages, prompt and instructions
(a) Primary passage	The item uses a passage, 'The Railroad Runs to Canada', from *Harriet Tubman: Conductor on the Underground Railroad*, by Ann Petry.
(b) Auxiliary passage	The item uses a short paper, 'Seven Qualities of a Good Leader', by Barbara White.
(c) Prompt	The item includes a one-page prompt with background, writing instructions, a prompt question, and instructions for both the body and the conclusion of the desired response.
(d) Instructions	The instructions come in the form of a four-page document with heading 'Pathway Project Reading and Writing Assessment'.

Note. HT = Harriet Tubman.

5.3.2.3 Scoring criteria and level definitions

The RDF criteria require explicit scoring (or evaluative) criteria, level descriptors, and level definitions. The initial RDF for Scenario 3 has the following seven scoring criteria, each with their own level descriptors and definitions. Points awarded for each narrative element category are either 1 or (in five of seven cases) 2. The total points across all evaluative criteria, limited by each evaluative category maximum value, are added together for a final score on a 0–12-point scale. This final raw score is transformed to a final scaled score of 0 to 6 points for comparability to both the C+E rubric and the holistic rubric for these item responses.

Table 5-27. Scenario 3, Phase 1: Scoring (Evaluative) Criteria

3. Scoring criteria	HT item: RDF A–G rubric
(a) Claim or 'a' subscore	Does the response identify a key quality of leadership that is responsible for HT's success? (2 points)
(b) Reasoning or 'b' subscore	Does the response describe why the chosen quality of leadership was selected? Does it include reasoning from observations? (2 points)

3. Scoring criteria	HT item: RDF A–G rubric
(c) Similar reactions or ‘c’ subscore	Does the response identify common reactions between HT and her followers to life-threatening situations? (2 points)
(d) Differing reactions or ‘d’ subscore	How were HT and her followers’ reactions different to life-threatening situations? (2 points)
(e) How is HT like her followers?	What does HT have in common with her followers? (max 1 point)
(f) How is HT different from her followers	What differences allow HT to emerge as a leader? (2 points)
(g) Lesson subscore	Does the story provide a lesson from HTs acts of courage? (max 1 point for lessons)

Note. HT = Harriet Tubman; RDF = rubric design framework.

Table 5-28. Scenario 3, Phase 1: Level Description and Quality Level Definition

3. Level descriptors	Quality level definitions (A–G)
(a) Subscore a, 2 points	(a) Identification of key trait. Additional points for clarity, or reasoning and context
(b) Subscore b, 2 points	(b) Reasoning and discussion supporting the claim
(c) Subscore c, 2 points	(c) HT’s similarity in response to life-threatening situations
(d) Subscore d, 2 points	(d) HT’s differing response to life-threatening situations
(e) Subscore e, 1 point	(e) What does HT have in common with her followers?
(f) Subscore f, 2 points	(f) How is HT different from her followers, in ways that help her succeed?
(g) Subscore g, 1 point	(g) Lessons learned from HT’s acts of courage

Note. HT = Harriet Tubman.

5.3.2.4 Subscale score calculation formula

Each narrative element subscore for each category is scored as 1 or 2 points, as indicated above. Any element requirement that is squarely and fully addressed is scored at the maximum, with the five 2-point maximum subscores set to 0 if not addressed, and 1 if addressed in a minimal or partial matter.

5.3.2.5 Final raw score formula

The raw score formula sums the subscores (capped by the subscore maximum allowed value) resulting in scores between 0 and 12.

5.3.2.6 Score scaling formula, and descriptors

The final score scaling formula divides the 0–12 scale score by two, rounding up to a whole number and using the whole number result, giving a final scaled score of 0 to 6. The scores then rest on the same scale as the original 6-point scale. The six quality level descriptors below reflect each score point's representation of the coverage of the seven possible narrative elements on a continuum. These numeric scores are also on the same scale as the C+E rubric in Scenario 2, offering additional perspective on the scores provided by the differing rubric criteria and allowing an examination of the relationship between success in following the narrative element suggestions and other evaluations such as the strength of argumentation.

Table 5-29. Scenario 3, Phase 1: Final Score Descriptor

Final score	Final score descriptor
6	Robust coverage of narrative element categories
5	Commendable narrative element coverage
4	Substantial narrative element coverage (at least four types)
3	Some narrative element coverage (at least three types)
2	Little narrative element coverage (at least two types)
1	A single narrative element coverage type
0	Nonresponsive

5.3.2.7 Score process, strategy, and design

This item scoring process is essentially the same as used in Scenario 2, albeit with different subscore evaluative criteria and quality level definitions.

Table 5-30. Scenario 3, Phase 1: Scoring Process

7. Scoring process	Strategy and design
(a) How scoring decisions are recorded	Scoring decisions for identifying specific narrative elements a–g are recorded in association with a specific sentences in the response. These indicate the number of points awarded: 1 or 2, depending on the category and strength of coverage provided.

7. Scoring process	Strategy and design
(b) How scoring data are used to produce a score report	With the association of all narrative element point awards to specific sentences captured during scoring, scoring reports can show strength of coverage by including supporting detail that indicates where specific coverage was identified for each narrative element. With these data, feedback for the student can also identify which narrative elements were neglected or insufficiently addressed.
(c) How meaningful feedback is produced	Giving specific, in-response feedback showing which narrative elements were used or neglected, students can assess on their own their performance with respect to the instructions specified.

5.3.2.8 Scoring process implementation

The scoring implementation process for Scenario 3 is the same one used in Scenario 2, which is described in Section 5.2.2.8.

5.3.2.9 Format and content of score reports

The scoring reports created during this study for Scenario 3, Phase 1, are the same as those used in Scenario 2, except that there are seven subscores per item response rather than just two. The reports show per category or evaluative criteria scores per sentence and response, with the maximums, and provide the totals and a final scaled score descriptor. In practice this means that each sentence (row) in the scoring grid is preceded by seven columns for scoring the presence of narrative elements a–g, rather than the two columns of Scenario 2, which accommodate subscores for claims and evidence. Format and content, and how such sentence-to-rubric associations can provide the basis for feedback, are addressed in Section 5.2.2.9 and the final chapter of this thesis.

5.3.2.10 Exemplars

Exemplars were not identified for Scenario 3 during this study, although a library of scored responses at various score points would provide a useful starting point for establishing a set of exemplars for this item prior to large-scale deployment.

5.3.3 Holistic scoring results

Scenario 3 compares the rater performance for scoring these item responses with each other while using the RDF-based 'seven narrative elements' rubric, and with the interrater performance for scorers using the holistic (and more general quality of writing) rubric of the original item, which was also used as the baseline of comparison in Scenario 2, as shown in Table 5-20). This IRR for scoring the HT item with the original holistic scoring by two raters is repeated for convenience as Table 5-31.

Table 5-31. Scenario 3, Phase 1: Holistic H1 vs. H2 Comparison

HT Item - h1 v h2 - Original Holistic Rubric -- Initial 40 Item Responses								
		criteria	total					
Agreement / Accuracy:		22	40	55%				
Adjacent Agreement:		40	40	100%				
Kappa with Quadratic Weighting			.95 Confidence Interval					
Observed Kappa		Std. Error	Lower Lmt | Upper Lmt					
0.9080		n/a	n/a | n/a					
H1/H2	**0**	**1**	**2**	**3**	**4**	**5**	**6**	
0								
1		2	2					4
2		2	4	1				7
3			3	4	2			9
4				2	3			5
5					2	6	2	10
6						2	3	5
		4	9	7	7	8	5	**40**
0.9796	maximum possible quadratic-weighted kappa, given the observed marginal frequencies							
0.9269	observed as proportion of maximum possible							

Note. H1/H2 = human raters; HT = Harriet Tubman.

5.3.4 Initial RDF scoring results

Using the RDF rubric defined for Scenario 3 in Section 5.3.2, two raters scored the same 40-item responses that had been scored with the holistic rubric. The results of comparing the two raters' scores using this rubric are shown in the confusion matrix

Table 5-32. The RDF scoring for HT with the A–G narrative elements rubric was less in agreement than those obtained by the scorers with the holistic rubric, with lower levels of exact and adjacent agreement (50% and 85%, as compared to 55% and 100%, respectively). The RDF scorer comparison showed a reasonably high level of overall agreement (QWK of 0.8476), though lower than the very high QWK observed for holistic scorer agreement of 0.9269. The RDF IRR measure was somewhat lower than for the holistic score but still above thresholds typically used to validate operational items for production use in high-stakes testing (e.g., a QWK exceeding 0.70, per Williamson, Xi and Breyer, 2012, p. 7).

In comparing the scores using the two rubrics across their range of results, I also note that in six of 40 items the final scaled scores of the two RDF scorers differed by more than 1 point, with four of these differing by 2 points. For the two remaining differentially scored items in this outlier group, in one instance the scorers differed by 3 points and the other the difference was 4 points (on the 0 to 6 scale). The analysis that follows considers these differences generally, with a focus on the greatest differences or common themes across the smaller differences.

Table 5-32. Scenario 3, Phase 1: RDF Scorer Comparison

Scenario 3 - h1 v h2 - Initial RDF Rubric -- 40 Initial Item Responses								
			criteria	total				
Agreement / Accuracy:			20	40	50%			
Adjacent Agreement:			34	40	85%			
Kappa with Quadratic Weighting				.95 Confidence Interval				
Observed Kappa		Std. Error		Lower Lmt | Upper Lmt				
0.8476		n/a		n/a | n/a				
H1/H2	**0**	**1**	**2**	**3**	**4**	**5**	**6**	
0	7							7
1	1							1
2		1	2					3
3			3		1			4
4			3		2	3		8
5		1	1			6	1	9
6					1	4	3	8
	8	2	9	0	4	13	4	**40**
0.9030	maximum possible quadratic-weighted kappa, given the observed marginal frequencies							
0.9386	observed as proportion of maximum possible							

Note. H1/H2 = human scorers; RDF = rubric design framework.

5.3.5 Phase 1 holistic versus RDF rubric results side by side

The confusion matrix in Table 5-32 reveals six score pairs outside the adjacent agreement zone of scores running down and across the diagonal of the table. These six outliers include scores with a 3- and a 4-point difference along with four instances of 2-point differences. A summary description of the comparative results for the holistic and RDF-based rubric scoring for this first phase of Scenario 3 is shown in Table 5-33. These results are analysed in the paragraphs that follow.

The overall shape of the score distributions for the holistic rubric and the RDF rubric were more similar in this scenario than in the others. As shown in Figure 5-10, the distribution of scores across the final score values is quite similar save for the difference of the zero score values assigned by the RDF scorers. As with the other scenarios, the RDF scores include a greater number of 0 score results, for similar

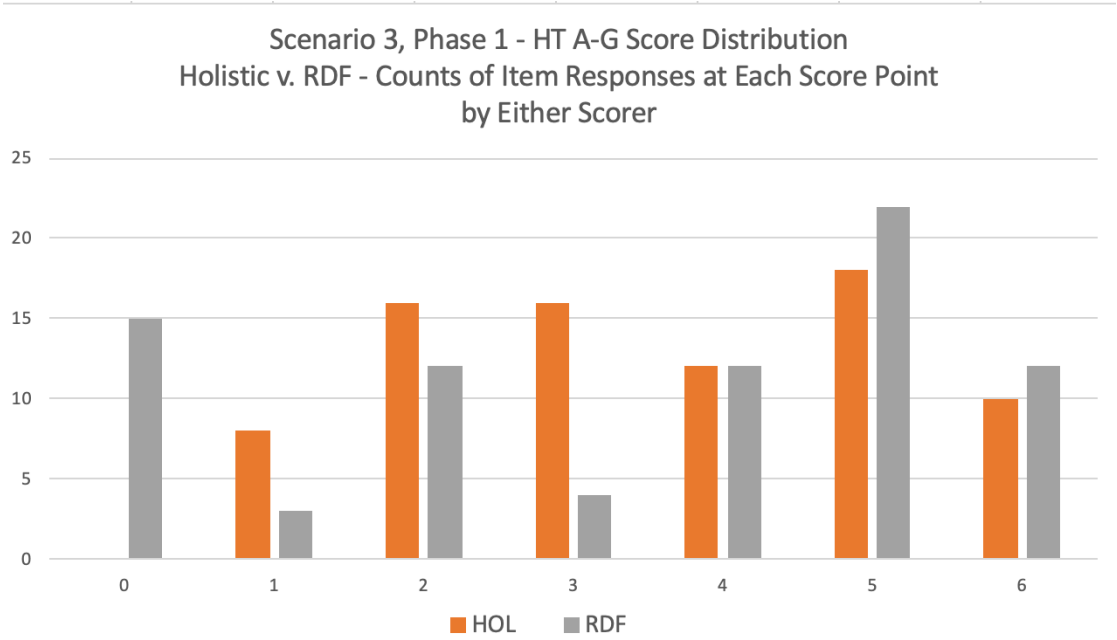
reasons, although in this case the number of 0 scores was less than half of those found in Scenario 2. The limited number of evaluative criteria in Scenario 2 versus Scenario 3, two subscores versus seven, and their slightly more general nature, are factors that are examined in the analysis of the scoring.

Table 5-33. Scenario 3, Phase 1: Holistic vs. RDF Scoring Comparison

Interrater (H1 vs. H2) and distribution comparisons	Holistic rubric	RDF rubric
Number of item responses	40	40
Accuracy	55%	50%
Adjacent agreement	100%	85%
QWK	0.9080	0.8476
QWK standard error	n/a	n/a
Average score	3.58	3.36
Standard deviation	1.57	2.12
Median score	3.5	4

Note. H1/H2 = human scorers; QWK = quadratic weighted kappa; RDF = rubric design framework.

Figure 5-10. Scenario 3, Phase 1: Holistic vs. RDF Score Distribution Chart



Note. HT = Harriet Tubman; RDF = rubric design framework; HOL = holistic.

5.3.6 Scoring analysis

The analysis of the scoring here will focus on the most divergent scoring examples and consider the sources of the different scoring decisions by different scorers. After a review of specific instances of scoring differences and describing them in general terms, the following section will identify and define areas where the rubric itself, or the rubric and the RDF, can be enhanced to improve scoring outcomes

On review of the most divergent score pairs for item responses in the development exercise, a number of issues were identified for analysis here:

- The most divergent set of scores, the pair with a 4-point difference and the pair with a 3-point difference on a 6-point scale, were reviewed to consider the potential source of such a large difference of opinion (Item Responses 13626 and 13699, respectively).

- Two other item responses among the group of 2-point differences were selected for review, as they reveal common themes in the sources of difference seen in these and other items in the group for this scenario (Item Responses 13613 and 13650).

Each of these four item responses and the issue they raise are analysed below.

5.3.6.1 Nonresponsive responses

Item responses for an AW writing challenge should, in most situations, be deemed unscorable if they fail to make an argument. Some scorers will identify portions of the text that could serve as an argument and find partial credit. Writing samples can be scored for a great variety of purposes, but when the purpose is to develop argumentation skills or to gain practice making a claim supported by evidence, responses that fail even to attempt the effort in most cases cannot be usefully scored with the intended rubric. An example, Item Response 13626, is presented in Figure 5-11 below. It received a partial credit for selected narrative elements based on the content of certain individual sentences devoid of context. Many individual sentences lifted from the passage or quoted did not present the content—that both HT and her followers were tired and hungry, for example—within a context of comparing or

contrasting HT with the followers. For more consistent and meaningful scoring, the rubric specification at the highest level should establish the parameters for a scorable essay as clearly as possible.

Figure 5-11. Item Response 13626: Off-Topic Response

Item Response 13626, complete

Do you ever wonder why we have this freedom today. "The Railroad to Canada" from Harriet Tubman: Conductor or the underground Railroad a biogrophy by Ann Petry.
Harriet Tubman wanted to run off slaves when she heard stories about them and how they captured them to be slaves. some examples of Harriet Tubman was that she was allways trustworthy of what she says and how asky she is when she risk her life to save the slaves.
In 1851 she led eleven runaway slaves all the way to Canada. As they walked along, she told them stories of her own first fight she kept painting vivid word pictures of what it would be like to be free. If they knew they got caught, the eleven runaways would be whipped and sold South, but she-she would probaly be hanged.
That night they reached the next stop she made runaways take shelter behind trees at the edge before she knocked at the door. They spent the night in the warm kitchen. They slept and when they left, it was with reluctance. Harriet had found it hard to leave the warmth and friendly.
The next day, she told them about Frederick Douglas and how he escaped of being salve. But they had been tired too long, hungry too long, afraid too long, footsore too long. She carried a gun with her on these tips, she had never used it -exept as a threat. As she aimed it and she experienced a feeling of guilt and rememberd she had prayed for the death of Edward Broads.
Finally, she gave the impression of being short, indomitable woman who could never be fedeated. Suddently they fell asleep in the woods. She was leading them into freedom, and so they waited until she was ready to go on. They stopped at Philadelphia and thought it was safe. They lived in whatever part of town they chose and sent their children to the schools.
Harriet Tubman was the first woman to lead on armed expedition in war to helped to liberate more than 700 slaves. She now works for the union Army to help lead the fight for the abolition of slavery.

5.3.6.2 Garbled text

Item responses sometimes contain garbled text, either due to limited facility with language on the part of students or as a result of technology or other issues. As a result, different scorers may take different approaches to text whose meaning is not immediately apparent. In the case of Item Response 13699, the final 10 sentences or sentence-like structures of text—nearly two-thirds of the response—were a particular challenge for scorers. One scorer saw only gibberish and scored a bit of credit for

information in the first portion of the response only, whereas another scorer seemed to read meaning into generous sections of the final portion of the response. The result was a significant 3-point difference in the scores (on a 6-point scale), with one scorer awarding an overall score of 5 and the other a 2. A review of the final portion of this response is presented in Figure 5-12. A careful reading suggests that it might be problematic to justify scoring any portion of this text as addressing one of the key narrative elements identified in the rubric. Again, the best course would be to mark such work unscorable and address the underlying issues with the writing rather than attempt to measure it against some standard with the intention of critiquing specific aspects of CT or argumentation. Figuring out the intention or meaning of 'she would fall asleep out of hoe hern'[9] is time spent on speculation rather than educating the student.

Figure 5-12. Item Response 13699: Garbled Text

Item Response 13699, final sentences
…Her dedication comes to show many time in this article. like when she was left in the cold with 11 other lives to worry about what do you do? Do you return the slaves back and risk being caught going back or do you keep going? Hungry, cold, tired and with morale low her dedication is what let her manage to get going she could've easily returned them and simply left. But she knew that she had a mission to do and that's what she was gonna do. Do you think that if it was just any person they hodctve risked their health and their life the way harriet tubman did. Its that one special and vital triat that they needed to make up a leader. Another example is when she would fall asleep out of hoe hern. It talks about how she fell asleep because her body gave out on her the only weary she stopped in when her body finally decided to give out on her. how hard do you have to work to have that happen to you the dedication that is needed to ignore your body of pain should be immense, she stuck through it because of her necescitie to finish what she started. Thats how dedication helped her finish what she needed to.

5.3.6.3 Implicit contrast

A recurring theme among scoring differences reviewed in this rubric development phase is the degree to which argumentation, reasoning, or the recitation of evidence can be reasonably inferred when the purpose of an observation is not explicitly stated in a response. Among the responses in Scenario 3, there were many instances where a

[9] Possibly 'exhaustion'.

narrative element would identify common characteristics between HT and her followers or identify contrasts between them, but the reason for the observation was not explicitly stated. In some cases, a rater would recognise the juxtaposition of parallel or opposed observations as making a point, even when not called out explicitly as reflecting argumentation or reasoning. At times these challenges were exacerbated by issues with spelling, grammar, or other problems that hid the observation's intent or meaning. An example that embodies both elements (lack of explicit reasoning and confusing syntax and spelling) can be seen in Item Response 13613, as shown in Figure 5-13.

Scanning, a grader could miss that the student meant to write 'She would also inspire [others] … by telling them lies [stories] just to motivate them not to stop'. During the development phase, one scorer did not award points for the observations of traits and characteristics that set HT apart from her followers, perhaps due to the confounding effects of *whold*, *insior*, and *motain*; another scorer was more generous in this and similar instances, creating a 2-point difference between the scorers in the final scaled score of the response.

Figure 5-13. Item Response 13613: Implicit Contrast, Errors

Item Response 13613, Sentences 10-11

She whold also insior By telling them Lies just to motain them not to stop or give up to keep on going. I also think that it takes dedication to memerise the was to go how long it will tack to got to a sertine place how she know where she had to stope.

Another example, Item Response 13650, illustrates the ambiguity that can arise from trying to draw inferences from a response. Item Response 13650 included several sentences that some scorers saw as highlighting differences between HT and the others that allowed her to be successful, so meeting the requirements for narrative element F of the rubric for Scenario 3; other scorers did not associate them with narrative element F because there was no explicit language such as 'this shows how she is different from the others'. Among the sentences counted by some scorers but not others as meeting the requirements of narrative element F (i.e., 'identifying differences between HT and the followers that enabled her to be a leader') were these four (underlined in Figure 5-14):

- ‘Tubman had promised them food and shelter so she didn't let them down.’
- ‘This was important because it proved that Harriet was trustworthy and that the she believed in herself.’
- ‘Fortunately that didn't happen because they had come to trust her implicitly.’
- ‘Harriet knew that she was trustworthy and she believed.’

Figure 5-14. Item Response 13650: Implicit Contrast

Item Response 13650. Complete. [Four underlined sentences.]

Have you ever wondered what the world would be like without leaders? Well it probably wouldn’t be in good shape. There are several qualities that make leaders guide others. In the biography of Harriet Tubman she had many leadership skills but the one that was most essential was trustworthiness. If the people are going to follow the leader they must trust them first and that’s what Harriet Tubman had. The slaves trusted her in bringing them north.

While taking eleven slaves up North toward Canada Harriet was determined to give these slaves freedom. She had never been in Canada so she was half afraid and always looking back. Harriet knew that is they were to get caught the eleven runaways would be whipped and returned to the Maryland plantations but she would be hanged. When they reached a farmhouse the owner didn’t let Harriet inside because he had been searched and the place was no longer safe. Tubman had promised them food and shelter so she didn’t let them down. She led them to next stop. A farm that belonged to a German. They were let in where they ate and spelt all night until dusk the next day. This was important because it proved that Harriet was trustworthy and that the she believed in herself.

When they started walking again it had been too long. They were tired hungry, afraid, and footsore. Until one man cried out loud, “Let me go back. It is better to be a slave than to suffer like this is order to be free. Harriet couldn’t let him go back because everyone that helped her trusted her in not exposing them. If the man went back the plantation owners would beat him until he spoke. “We got to go free or die. And freedom not bought with dust.

It was obvious that she was tired as well so she fell asleep knowing that the runaways could grab her gun and go their own way. Fortunately that didn’t happen because they had come to trust her implicitly. They sat down and waited paitently. Until she awoken.

Harriet knew that she was trustworthy and she believed that she could do anything as long as the people supported her. From her story we can learn that by telling the truth people will trust you.

5.3.7 Scenario 3 RDF rubric adjustments

The analysis of the examples above suggests the following revision to the Scenario 3 rubric and has implications for the RDF itself as summarised below.

5.3.7.1 Nonresponsive or off-topic responses

Scoring item responses with an RDF rubric is first informed by the definition of what is being scored in the high-level rubric definition, RDF Element 1. Sub-element 'd' of Element 1 addresses the 'how scored' topic at a general and evaluative quality level, where it is appropriate to identify how scoring should address off-topic and nonresponsive item responses. Rather than devote energy and effort to figure out how best to deal with such responses on a case-by-case basis, this rubric (and other CT rubrics) should recognise explicitly that formulating AW and CT responses is a skill that builds on fundamental skills and knowledge related the standards and conventions of language usage. Just as CT requires sufficient knowledge of a topic to think about it critically, written CR responses that need to express claims, cite evidence, explain reasoning, and develop argumentation require foundational language skills that, if not present, limit the utility of an instrument that requires these skills as a prerequisite.

In most cases for CT and AW assessment, and in all the rubrics in this study, the 'how scored' section of the rubric should be augmented to note that item responses that do not address the demands of the prompt in fundamental ways should be marked as 'not scored' with an explanation that might include the rationale for the decision. Too many fundamental errors in grammar, syntax, spelling, or other mechanical issues might rule out scoring; a substantial portion of a response that is not legible, does not read as the language of the exam (in the case of this study, English), or is unrecognisable as standard English text warrants a similar treatment. And just as importantly, when an assessment that defines a task with specificity is not addressed or acknowledged by the response (e.g., no claim is made when one is demanded, no position is taken when one side or another of a proposition is to be argued), such item responses also need to be identified as nonresponsive, illegible, or off topic; they should not be, and indeed cannot reasonably be, scored by the rubric.

An example of how the HT A–G narrative elements rubric for Scenario 3 has been adjusted to address off-topic or nonresponsive item responses for the testing phase of Scenario 3 is illustrated by the additional paragraph added to the rubric definition (previously shown in Table 5-25) as shown in Table 5-34 below.

Table 5-34. HT A–G Rubric: RDF Element 1 Adjustments

1. Rubric definition	HT RDF narrative elements (A–G)
(d) How scored (First paragraph added)	Scorers will not score off-topic, nonresponsive answers or responses written in English that in their judgement is not clear and correct enough to convey their intent and their understanding of the task. That is to say, if the body of the essay does not address the claim of a most important trait to enable HT to succeed or describe the factors that hinder or enable her success, it need not be scored. Retelling the story in the passage is an example of a nonresponsive item response. Scorers will assign points to content by identifying which of the seven content types are present in the response and score a 1 or 2 based on the number and depth of observations of each type. These point allocations will be attributed to individual sentences. For the expression of an idea that is composed of contributions from multiple sentences, the points will be associated with the first sentence of the group.

(Replacement to row (d) of rubric definition in Table 5-25 in Section 5.3.2.1)

To further address the idea that narrative elements defined in this rubric have more specific requirements than are articulated in Element 3 in either the scoring criteria or the quality level definitions, those items should also be augmented to more fully represent the intention of the rubric.

For example, the rubric for this item could have the quality level definitions for subscores c, d, e, and f define both the narrative element topic *and* that these narrative elements must also articulate specifically how the differences or similarities in question either add to HT's challenges or improve her likelihood of success. An example of this augmentation to the quality level definitions is shown in

Table 5-35 (changes in bold red) below.

Table 5-35. HT A–G Rubric RDF: Element 3 Adjustments

3. Level descriptors	Quality level definitions (A–G)
(a) Subscore a: 2 points	(a) Identification of key trait. Additional points for clarity or reasoning and context.
(b) Subscore b: 2 points	(b) Rationale/reasoning and discussion supporting the claim; why the identified trait is the most important.
(c) Subscore c: 2 points	(c) HT's similarity in response to life-threatening situations. **How this helps or hinders her success.**
(d) Subscore d: 2 points	(d) HT's differing response to life-threatening situations. **How this helps or hinders her success.**
(e) Subscore e: 1 point	(e) What does HT have in common with her followers? **How this helps or hinders her success.**
f) Subscore f: 2 points	(f) How is HT different from her followers, in ways that help her succeed?
(g) Subscore g: 1 point	(g) Lessons learned from HT's acts of courage
Additional notes (**added for Phase 2**)	Award more than one point where possible based on substantial or clear discussion, or one that includes extended remarks.

Note. HT = Harriet Tubman; RDF = rubric design framework.

5.3.7.2 Garbled text

Like nonresponsive item responses, item responses made up of garbled text should not be scored. Scorers are free to make inferences and allowances that reflect the anticipated writing and English abilities of the target audience, but the adjustments made to the rubric for this scenario above are sufficient guidance to address unreadable text.

5.3.7.3 Implicit reasoning or connections

After review and discussion with scorers, there was recognition and agreement that in everyday usage, some expressive techniques convey inferences with obvious intent, such as juxtaposing two contrasting traits or sets of facts to create contrast that is both obvious and a direct and reasonable inference. At the same time, they also recognised that scoring a nonresponsive paper, such as one that retold the story of the primary

passage rather than making a claim as required by the prompt, could lead to more confusion about how to apply the rubric to something it was not designed to measure.

Scorers also advocated that, with complete quality level definitions, the scorer should not need to make complex judgement calls about the intent of an examinee's writing. If an item response needs to explicitly identify a bit of evidence or articulate a claim with specific precision, or explicitly cite reasoning to connect an observation to an inference or a claim, then this should be made clear in the rubric; otherwise, scorers should apply the same standards when making inferences from item responses that they would in other academic writing, given the medium, the environment, and the context. Academic writing in general is usually explicit. Making an argument in an academic paper does not rely on the reader to fill in missing inferences or gaps in reasoning or logic, even when it requires commonly known or understood knowledge. Consequently, I made no further adjustments to the rubric for this scenario to address this issue. The rule, simply put, was 'Score based on what is written, not what you believe was intended'.

Chapter 6 Rubric Design Framework Testing (Phase 2)

This testing phase of the work used the revised rubrics as updated in the development phase to score a greater number of samples with two scorers. As with the development stage, the item responses represent the full range of score values as produced by the original holistic rubric, and the new RDF rubric-based scores are measured and compared in terms of IRR to the two human scores assigned by the original raters using the holistic rubric. In each scenario the results of the testing phase scoring will be compared with the original holistic scoring for those items, and to the results overall from the development stage RDF scoring, to assess the nature and magnitude of any improvement.

6.1 Scenario 1 – WH Claim and Evidence Rubric Testing Phase

In this testing phase for Scenario 1, the updated claim and evidence rubric developed in Chapter 5 and revised in Section 5.1.7 was applied by two raters to an additional 120 WH item responses that averaged less than eight sentences each. The scoring for this scenario includes (a) a claim that recognises and describes an analogy implied in the original item materials, or some aspects of it; and (b) the presentation of evidence to support the analogy (or at least the portion of it that was recognised).

6.1.1 Test phase holistic scoring baseline

To test the updated C+E RDF rubric for this scenario, a larger random sample of 120 item responses was selected for analysis of the original holistic scoring to serve as a baseline for comparison to the RDF scoring for these same items. This section describes this new population-specific scoring in terms of IRR and score distribution.

As in the development stage, the proportion of items at each score point between 1 and 3 was roughly equal, with few items having a 0 score. The population of holistic scores for the item responses in the test sample is similar in makeup to those in the development group, as is clear from the confusion matrix in Table 6-1 and the population comparison that follows in Table 6-2.

Table 6-1. Scenario 1, Phase 2: Holistic H1 vs. H2 Score Comparison

Scenario 1 Phase 2 - h1 v h2 - holistic rubric -- 120 responses					
		criteria	total		
Agreement / Accuracy:		78	120	65%	
Adjacent Agreement:		120	120	100%	
Kappa with Quadratic Weighting			.95 Confidence Interval		
	Obs. Kappa	Std. Error	Lower Lmt	Upper Lmt	
	0.6860	0.1176	0.4555	0.9165	
H1/H2	**0**	**1**	**2**	**3**	
0	2				2
1		27	12		39
2		7	36	10	53
3			13	13	26
	2	34	61	23	**120**
0.9402	maximum possible quadratic-weighted kappa, given the observed marginal frequencies				
0.7296	observed as proportion of maximum possible				

Note. H1/H2 = human scorers.

The summary information in Table 6-2 shows the similarity in results from holistic scoring for the populations of item response scores in the development and testing groups. While the accuracy in terms of exact match was greater in this group, the overall IRR for this group is quite similar to the development group population, as expected.

Table 6-2. Scenario 1: Holistic Scoring: Development vs. Test Groups

Interrater (H1 vs. H2) and distribution comparisons	Holistic rubric development group	Holistic rubric testing group
Number of item responses	40	120
Accuracy	50%	65%
Adjacent agreement	100%	100%
QWK	0.6537	0.6860
QWK standard error	0.1430	0.1143
Average score	1.93	1.87
Standard deviation	0.85	0.75
Median score	2	2

Note. H1/H2 = human raters; QWK = quadratic weighted kappa.

6.1.2 Changes to the Scenario 1 rubric

As described in the development phase results for Scenario 1, the rubric was revised to help resolve ambiguities in the quality level definitions in initial RDF formulation of the proposed rubric for this item. The changes were to simplify scoring for the partial recognition of the underlying analogy in the main passage and provide additional guidance to ensure that only evidence that was actually cited explicitly or by direct implication or reasoning would be counted as evidence. The latter condition was added to avoid counting content that was simply retelling the story from the main passage as citing evidence. These changes were designed to resolve ambiguities and improved the associated feedback that inappropriate scoring would imply. The revised rubric elements appear in Table 5-10 and Table 5-11.

6.1.3 Scoring additional items

The results of scoring the additional 120 items by two human scorers using the improved RDF rubric for the WH claim and evidence rubric are shown in Table 6-3.

Table 6-3. Scenario 1, Phase 2: RDF H1 vs. H2 Score Comparison

Scenario 1 Phase 2 - h1 v h2 - RDF rubric -- 120 responses					
		criteria	total		
Agreement / Accuracy:		82	120	68%	
Adjacent Agreement:		120	120	100%	
Kappa with Quadratic Weighting			.95 Confidence Interval		
	Obs. Kappa	Std. Error	Lower Lmt	Upper Lmt	
	0.8337	n/a	n/a	n/a	
H1/H2	**0**	**1**	**2**	**3**	
0	10	5			15
1	4	44	11		59
2		1	9	10	20
3			7	19	26
	14	50	27	29	**120**
0.9037	maximum possible quadratic-weighted kappa, given the observed marginal frequencies				
0.9225	observed as proportion of maximum possible				

Note. H1/H2 = human scorers; RDF = rubric design framework.

The greater number of examples yielded a similar but slightly improved IRR than did the development scoring work, as shown in the side-by-side comparison in Table 6-4.

Table 6-4. Scenario 1: RDF Scoring: Development vs. Test Groups

Interrater (H1 vs. H2) and distribution comparisons	RDF development group	RDF test group
Number of item responses	40	120
Accuracy	63%	68%
Adjacent agreement	93%	100%
QWK	0.7073	0.8337
QWK standard error	0.1302	n/a
Average score	1.4	1.53
Standard deviation	1.04	1.01
Median score	1	1

Note. H1/H2 = human scorers; QWK = quadratic weighted kappa; RDF = rubric design framework.

6.1.4 Analysis of the scoring results

To recap the data comparisons between both the holistic and RDF-based rubric scores, they are all presented side by side in Table 6-5 to help visualise the comparison across and between paired scorers.

Table 6-5. Scenario 1: Holistic and RDF, Development and Test Scoring Results

Interrater (H1 vs. H2) and distribution comparisons	Holistic		RDF	
	Development	Test	Development	Test
Number of item responses	40	120	40	120
Accuracy	50%	65%	63%	68%
Adjacent agreement	100%	100%	93%	100%
QWK	0.6537	0.6860	0.7073	0.8337
QWK standard error	0.1430	0.1143	0.1302	n/a
Average score	1.93	1.87	1.40	1.53
Standard deviation	0.85	0.75	1.04	1.01
Median score	2	2	1	1

Note. H1/H2 = human scorers; QWK = quadratic weighted kappa; RDF = rubric design framework.

The revised RDF rubric resulted in a significant increase in the overall agreement rates between the raters, with accuracy and adjacent agreement both increasing nearly 10%. Using the Landis and Koch (1977) classification system terms described in section 4.8, this change from 0.7073 to 0.8337 is significant and moves IRR from *substantial agreement* to *almost perfect agreement*.

The higher median score and average score for the revised RDF rubrics compared to the initial development phase scoring reflects that the scores were generally higher (including a big reduction in zero-scored responses) with the revised rubric. The slightly higher QWK and slightly lower standard deviation reflect a slight improvement in the clustering of the scores in that narrower higher range constrained by fewer zero scores. An examination of the data shows that the proportion of 0 scores awarded during development scoring was 22.5% (or 18 of 80, from Table 5-8) for the two scorers combined, which fell to 12.1% (or 29 of 240, from Table 6-3) during the testing phase.

It is also worth noting that in both stages the accuracy and QWK measures of IRR were higher for the RDF rubric than for the holistic rubric, and that they improved from the development to the testing state, with the best IRR, accuracy, and QWK results achieved for the revised final RDF-based rubric.

6.2 Scenario 2: HT Claim and Evidence Rubric Testing Phase

In this testing phase for Scenario 2, the C+E rubric developed in Chapter 5 in Section 5.2.7 was applied by two raters to an additional 80 HT item responses that averaged 19 sentences each. This rubric specifies scoring points based on (a) the identification of a specific character trait (from a set defined in the item) that most enabled HT (the protagonist of the primary passage) to succeed in her work; and (b) the presentation of evidence from the item materials to support this claim.

In this testing phase of Scenario 2, an additional 80 items were scored by two raters using the rubric as updated during the development phase. These responses, like those in the development phase, averaged 19 sentences and represented the full range of possible scores according to the holistic rubric scoring done on these item responses.

6.2.1 Test phase holistic scoring baseline

To test the updated HT C+E RDF rubric for this scenario, a larger random sample of 80 item responses was selected from the larger group of original holistically scored item responses. As in the development stage, the full range of holistic item scores was represented in this test phase sample, with a few items having a 0 score. The population of holistic scores for the item responses in the test sample is similar in makeup to those in the development stage case, as is clear from the confusion matrix for the holistic scoring by two scorers shown in Table 6-6. Also, the similarity between the two groups in terms of IRR and score distribution measures is shown in Table 6-7 in the comparison below the confusion matrix.

Table 6-6. Scenario 1: Holistic Scoring: Development vs. Test Groups

HT Item - h1 v h2 - Holistic Rubric -- Testing Phase 2								
			criteria	total				
Agreement / Accuracy:			40	80	50%			
Adjacent Agreement:			80	80	100%			
Kappa with Quadratic Weighting				.95 Confidence Interval				
Observed Kappa		Std. Error		Lower Lmt | Upper Lmt				
0.8989		n/a		n/a | n/a				
H1/H2	**0**	**1**	**2**	**3**	**4**	**5**	**6**	
0								
1		6	6					12
2		2	9	6				17
3			3	8	6			17
4				2	3	5		10
5					2	10	4	16
6						4	4	8
		8	18	16	11	19	8	**80**
n/a	maximum possible quadratic-weighted kappa, given the observed marginal frequencies							
n/a	observed as proportion of maximum possible							

Note. H1/H2 = human scorers.

Table 6-7. Scenario 2: Holistic Scoring: Development vs. Test Groups

Interrater (H1 vs. H2) and distribution comparisons	Holistic rubric development group	Holistic rubric testing group
Number of item responses	40	80
Accuracy	55%	50%
Adjacent agreement	100%	100%
QWK	0.9796	0.8989
QWK standard error	n/a	n/a
Average score	3.58	3.4
Standard deviation	1.57	1.57
Median score	3.5	3

Note. H1/H2 = human scorers; QWK = quadratic weighted kappa.

6.2.2 Changes to the Scenario 2 rubric

As described in the development phase results for Scenario 2, the rubric was revised to address ambiguities in the quality level definitions for the claim score (related to selection of traits outside those articulated in the item itself), and to clarify questions of when to score credit for citations of evidence when evidence (or at least what could be cited as evidence) is included in the item response but not actually used as evidence. These and other minor issues were addressed with additional descriptive text in the rubric for use in this testing and are shown in detail in Chapter 5, Section 5.2.7, including Table 5-23 and Table 5-24.

6.2.3 Scoring additional items

The results of scoring the additional 80 items by two human scorers using the improved RDF rubric for HT claim and evidence are shown in Table 6-8.

Table 6-8. Scenario 2, Phase 2: RDF H1 vs. H2 Score Comparison

Scenario 2 - h1 v h2 - RDF Rubric -- Testing Phase 2								
			criteria	total				
Agreement / Accuracy:			33	80	41%			
Adjacent Agreement:			74	80	93%			
Kappa with Quadratic Weighting				.95 Confidence Interval				
Observed Kappa		Std. Error		Lower Lmt | Upper Lmt				
0.8988		n/a		n/a | n/a				
H1/H2	**0**	**1**	**2**	**3**	**4**	**5**	**6**	
0	14	6	1					21
1	4	1						5
2	1	7	4	2	1			15
3	1		5	3	2			11
4			1	2	3	5	1	12
5					1	3	7	11
6							5	5
	20	14	11	7	7	8	13	**80**
0.9364	maximum possible quadratic-weighted kappa, given the observed marginal frequencies							
0.9598	observed as proportion of maximum possible							

Note. H1/H2 = human scorers; RDF = rubric design framework.

The greater number of examples yielded an IRR similar to but slightly improved over the development scoring work, as shown in the side-by-side comparison in Table 6-9.

Table 6-9. Scenario 2, HT Item, C+E Rubric: RDF Scoring: Development vs. Test Groups

Interrater (H1 vs. H2) and distribution comparisons	RDF rubric development group	RDF rubric testing group
Number of item responses	40	80
Accuracy	40%	41%
Adjacent agreement	78%	93%
QWK	0.7919	0.8988
QWK standard error	n/a	n/a
Average score	2.86	2.58
Standard deviation	2.10	2.09
Median score	3	2

Note. H1/H2 = human scorers; RDF = rubric design framework; QWK = quadratic weighted kappa.

6.2.4 Analysis of the scoring results

To recap the data comparisons between both the holistic and RDF-based rubric scores, they are all presented side by side in Table 6-10 to help visualise the comparison across and between paired scorers.

Table 6-10. Scenario 2: Holistic and RDF, Developmental and Test Scoring Results

Interrater (H1 vs. H2) and distribution comparisons	Holistic		RDF	
	Developmental	Test	Developmental	Test
Number of item responses	40	80	40	80
Accuracy	55%	50%	40%	41%
Adjacent agreement	100%	100%	78%	93%
QWK	0.9796	0.8989	0.7919	0.8988
QWK standard error	n/a	n/a	n/a	n/a
Average score	3.575	3.4	2.9	2.575
Standard deviation	1.57	1.57	2.10	2.09
Median score	3.5	3	3	2

Note. H1/H2 = human scorers; QWK = quadratic weighted kappa; RDF = rubric design framework.

The IRR for the holistic HT scoring measures was extremely high, reflecting the close agreement between raters across the 6-point scale and lower probability of chance agreement with the six-point (as compared to Scenario 1's four-point) rating scale. The IRR measures for the RDF rubric were themselves very strong, and as seen in Scenario 1, the final QWK measure for IRR moved from *substantial agreement* to *almost perfect agreement* by improving from 0.7919 to 0.8988. As also seen in Scenario 1, the improvement in the IRR scoring performance on the RDF rubric came with a further decline of average scores, indicating that additional rigour in the rubric and scoring might have achieved higher agreement in part by disallowing what might have been marginal point awards under the initial development form of the RDF rubric. And while the final IRR is essentially the same as achieved with the holistic scoring, the finer grained decisions and the capture of associations between score

points and individual sentences that enables detailed scoring feedback was achieved without any real degradation in the reliability of the scoring.

6.3 Scenario 3: HT A–G Narrative Elements Rubric Testing Phase

In the testing phase for Scenario 3, the 'A–G Narrative Elements' rubric developed in Chapter 5, Section 5.3.7, and refined in Phase 1 was applied by two raters to an additional 80 HT item responses—the same HT item responses used to test the rubric in Scenario 2. Using the same item responses means comparing the scoring performance of the raters against each other as well as with the scoring performance of the same baseline holistic scores used in Scenario 2. An implication of using the same item responses with two rubrics is that it could allow the data to be used (in a study outside the scope of this one) to consider the effect of the narrative elements aspect of the pedagogy on the C+E scores and the effect of their correlation with the holistic writing scores.

6.3.1 Test phase holistic scoring baseline

The holistic scoring for this scenario is the same as used in Scenario 2; those scores are reproduced in Table 6-11 from Section 6.1.1 for convenience.

Table 6-11. Scenario 3, Phase 2: Holistic H1 vs. H2 Comparison

HT Item - h1 v h2 - Holistic Rubric -- Testing Phase 2								
			criteria	total				
Agreement / Accuracy:			40	80	50%			
Adjacent Agreement:			80	80	100%			
Kappa with Quadratic Weighting				.95 Confidence Interval				
Observed Kappa		Std. Error		Lower Lmt | Upper Lmt				
0.8989		n/a		n/a | n/a				
H1/H2	**0**	**1**	**2**	**3**	**4**	**5**	**6**	
0								
1		6	6					12
2		2	9	6				17
3			3	8	6			17
4				2	3	5		10
5					2	10	4	16
6						4	4	8
		8	18	16	11	19	8	**80**
n/a	maximum possible quadratic-weighted kappa, given the observed marginal frequencies							
n/a	observed as proportion of maximum possible							

Note. H1/H2 = human scorers; HT = Harriet Tubman.

Also repeated for the holistic scoring of HT item responses is the comparison in Table 6-12 of the holistic scoring results from the development and this testing phase of Scenario 3.

Table 6-12. Scenario 3, Phase 2: HT Holistic H1 vs. H2 Score Comparison

Interrater (H1 vs. H2) and distribution comparisons	Holistic rubric development group	Holistic rubric testing group
Number of item responses	40	80
Accuracy	55%	50%
Adjacent agreement	100%	100%
QWK	0.9796	0.8989
QWK standard error	n/a	n/a
Average score	3.575	3.4
Standard deviation	1.57	1.57
Median score	3.5	3

Note. H1/H2 = human scorers; HT = Harriet Tubman; QWK = quadratic weighted kappa.

6.3.2 Changes to the Scenario 3 rubric

The Scenario 3 rubric development and analysis work of Chapter 5 found that nonresponsive or off-topic responses, garbled text, and an over-reliance on implicit contrasts and other inferences by scorers contributed to some of the most divergent scores (in Section 5.3.6), leading to adjustments to the rubric as shown in detail in Chapter 5, Section 5.2.7, including Table 5-23 and Table 5-24. Nonresponsive and unintelligible items were scored as zero, and graders were instructed to grade evidence that was cited for that purpose or when the implication was clear and direct (e.g., immediately before or after a relevant observation or implication).

6.3.3 Scoring additional items

The results of scoring 80 additional item responses by two scorers using the improved RDF HT narrative elements rubric are shown in Table 6-13.

Table 6-13. Scenario 3, Phase 2: RDF H1 vs. H2 Score Comparison

Scenario 3 - h1 v h2 - Revised HT A-G RDF Rubric -- 80 Item Responses								
			criteria	total				
Agreement / Accuracy:			41	80	51%			
Adjacent Agreement:			77	80	96%			
Kappa with Quadratic Weighting				.95 Confidence Interval				
Observed Kappa		Std. Error		Lower Lmt | Upper Lmt				
0.9201		n/a		n/a | n/a				
H1/H2	**0**	**1**	**2**	**3**	**4**	**5**	**6**	
0	4	3						7
1		5	2					7
2		2	7	3				12
3		1	1	2	3			7
4			1	3	6	3	1	14
5					3	6	7	16
6						6	11	17
	4	11	11	8	12	15	19	**80**
0.9833	maximum possible quadratic-weighted kappa, given the observed marginal frequencies							
0.9357	observed as proportion of maximum possible							

Note. H1/H2 = human scorers; HT = Harriet Tubman; RDF = rubric design framework.

Table 6-14 shows that with a greater number of examples and improved rater guidance, the RDF IRR performance improved on every measure of agreement from the development phase, as shown in the side-by-side comparison.

Table 6-14. Scenario 3: HT Item, A–GRDF Rubric Scoring Development vs. Test Group

Interrater (H1 vs. H2) and distribution comparisons	RDF rubric development group	RDF rubric testing group
Number of item responses	40	80
Accuracy	50%	51%
Adjacent agreement	85%	96%
QWK	0.8476	0.9201
QWK standard error	n/a	n/a
Average score	3.36	3.65
Standard deviation	2.12	1.94
Median score	4	4

Note. H1/H2 = human raters; HT = Harriet Tubman; QWK = quadratic weighted kappa; RDF = rubric design framework.

6.3.4 Analysis of the scoring results

To recap the data comparisons between the holistic and RDF-based rubric scores for this scenario, they are all presented side by side in Table 6.15 to help visualise the comparison across and between paired scores.

Table 6-15. Scenario 3: Development and Test Scoring Results

Interrater (H1 vs. H2) and distribution comparisons	Holistic		RDF	
	Development	Test	Development	Test
Number of item responses	40	80	40	80
Accuracy	55%	50%	40%	41%
Adjacent agreement	100%	100%	85%	96%
QWK	0.9796	0.8989	0.8476	0.9201
QWK standard error	n/a	n/a	n/a	n/a
Average score	3.575	3.4	3.36	3.65
Standard deviation	1.57	1.57	2.12	1.94
Median score	3.5	3	4	4

Note. H1/H2 = human raters; QWK = quadratic weighted kappa; RDF = rubric design framework.

The holistic scoring for the HT items in the scenario had very high interrater agreement rates. The IRR for RDF scoring, despite capturing significantly more detail in the rationale for the score and the specific response text that contributed to the score, was nearly as high and improved further with the improvements to the rubric and instruction. Indeed, all measures improved: Agreement was essentially flat at 40% and 41%, for the development and testing phases, respectively, but adjacent agreement and QWK improved significantly, from 85% to 96%, and from 0.8476 to 0.9201, respectively.

Chapter 7 Conclusions, Discussions and Closing Remarks

7.1 Introduction

This thesis has demonstrated that well-structured rubrics can enable nuanced and detailed scoring results without compromising the inter-rater reliability of scoring outcomes. These rubrics allow scoring to capture specific associations between response content and rubric quality level definitions, which then can associate useful feedback directly to student response content, helping students understand the scoring and learn from their score reports. Further, this same association provides assessment providers with defensible scores, and can achieve this without compromising scoring reliability (particularly with regard to IRR).

Taken together, these advantages for a new kind of CR scoring have the potential to convince more stakeholders that the assessments are valuable as learning experiences, that they provide a useful and authentic measure of what students know and can do, and that they favour the knowledge to be taught and measured over the 'test taking skills' measured in other forms of assessment that may be faster to deploy and far cheaper to score.

7.2 Findings

7.2.1 Feedback, education, justification, and reliability

The primary research question was whether a generalized and flexible rubric design framework could be used to create item-specific, content-centric rubrics that would result in scored responses that could provide useful feedback to students and teachers; nuanced scoring to enable assessment-as-learning; support explicit, defensible rationales for scoring outcomes; and do this with improved interrater reliability over scoring based on generic, holistic rubrics.

Two of the rubrics had evaluative criteria and quality level definitions tied to both claim and evidence scores. The RDF rubric structure's support for higher and lower degrees of quality and the association of these levels of quality with identifiable response elements were successful in underpinning scoring with an objective basis and led to scoring outcomes that could be documented and supported by those associations. In Scenario 1, scorers were consistently able to identify and support full credit for claims that explicitly identified the implicit analogy between Saeng and the

WH in the story, including their need to adapt and grow, from responses that reflected only determination or some other unidentified relationship between the WH and Saeng. In this way, claim scores for full credit could be explicitly supported, and scores with partial credit for recognition of a relationship or the theme of hard work and determination, could also be supported and differentiated, allowing score reports to be augmented with annotations noting both the deficiency and the source of what credits were awarded.

In another example of how effectively the rubric supported the research goals, the claim score in Scenario 2 was structured to recognise that a single most important trait was identified and to note that the trait was one of those specified in the auxiliary passage. Scores that identified and discussed a trait without a claim that it was most important, or selected a trait not included in the item materials, or identified more than one trait, could be given more or less credit as indicated in the rubric. This decision could then be supported in a score report with explanatory notes that highlighted the demands of the question, described how the response provided failed to meet the criteria, and provided exemplary examples of fully conforming claim statements from a library of examples made for that purpose, if an automated score reporting/feedback capability was included in a suitable assessment delivery platform.

In a third example from the evidence subscores that defined evaluative criteria, Scenario 1 had examples of using specific observations in the quality definitions for evidence, while Scenario 2 had evidence elements in its rubric that were more ‘kinds of observation’ categories. Scenario 1 anticipated specific WH traits called out in the definitions that could support the reasoning that the WH in Saeng’s new home was a variation on the WH she knew from her native land. This is in contrast to quality level definitions in Scenario 2 that described, for example, ‘ways in which HT differed from her followers in their reaction to life-threatening danger’.

These examples display elements of

- feedback, both of the sort that acknowledges successful performance and that identifies missing elements or incomplete analysis;

- education, in that the deficiencies can be explicitly addressed by feedback that informs and explains what a better answer would entail;

- justification, flowing from both the transparent and open scoring rationales and from the partial-credit-for-partially-correct variations; and

- improved reliability, or at least not significantly lower levels of reliability, where in every case the RDF rubric's score reporting detail and rationale is specific and detailed, against the collection-of-traits overall summary included in the holistic score. These improved markers of quality were present in every case, as seen in Table 7-1. Even when the broad generic measure had excellent IRR and the RDF scoring had lower but still very good IRR, many educators would likely prefer the approach with detailed feedback at the expense of a slightly lower but still highly acceptable IRR.

Table 7-1. Interrater Reliability: Summary of Holistic vs. RDF Results

Scenario	Holistic rubric	RDF rubric
Scenario 1: WH C+E		
Accuracy	65%	68%
Adjacent	100%	100%
QWK	0.6860	0.8337
Scenario 2: HT C+E		
Accuracy	50%	41%
Adjacent	100%	93%
QWK	0.8989	0.8988
Scenario 3: HT A–G		
Accuracy	50%	41%
Adjacent	100%	96%
QWK	0.8989	0.9201

Note. A–G = narrative elements rubric; C+E = claim + evidence; HT = Harriet Tubman; QWK = quadratic weighted kappa; RDF = rubric design framework; WH = Winter Hibiscus.

7.2.2 How did the RDF rubric facilitate scoring?

The secondary research questions asked, "Are there aspects of scoring with item-specific, content-centric rubrics that work well or that make scoring easier or more efficient?" My study showed that scoring with these rubrics could be both faster and easier than scoring holistically, as reflected by scoring times and comments from scorers. While we do not have a baseline of holistic scoring speeds for comparison, scoring times per item were uniformly low (less than two minutes per item) for rubrics with detailed quality level definitions (both claim and evidence rubrics). Scorers generally found the scoring 'easy'.

7.2.2.1 Clear quality level definitions led to faster and easier scoring

During scoring, the benefit of minimising the preferences and judgements of individual scorers, and particularly of limiting the effect of judgements not captured directly in the rubric on the quality level ratings for specific evaluative criteria, led to more consistent scoring. Explicit, well-defined and clear quality level definitions supported consistent discrimination by scorers, resulting for example in sub-scores for claim statements with very high IRR. Overall high levels of IRR reflect these strong consistencies in scoring results.

7.2.2.2 Scoring based on content detail led to faster and easier scoring

Quality level definitions that identified specific item content and clear criteria led to faster and easier scoring. Detailed evidence citations that indicated specific content as acceptable evidence led to fast and consistent scoring. Evidence that the WH in two places had evolved from a common ancestor could be specified in precisely those broad terms—or evidence of this could be called out in the rubric by crediting observations that the WH in both places had (a) similar shaped leaves, (b) similar leaf texture, (c) similar flower colour, or (d) similar fragrance. The latter approach is reflected in the rubrics and the high levels of IRR obtained for the scoring.

7.3 Implications

7.3.1 Benefits of RDF rubrics for scoring

My research has shown that the rubric design framework for critical thinking and argumentative writing developed in this study is well suited to the challenges of assessing critical thinking and argumentative writing.

1. They focus assessment on concrete decisions by examinees to address a challenge with specific kinds of information, providing an immediate and objective classification of responses into rough categories (e.g., with or without a claim, argument, evidence).

2. They provide separate quality level definitions for separate aspects of CT or AW to be measured, potentially providing more and better diagnostic information, and are especially useful for self-study and classroom settings.

3. They naturally break down the rater training required into discrete aspects of scoring.

4. They provide a consistent mechanism for scoring discrete aspects of CT or AW and for combining those scores into an overall score and into any sort of scale to be developed and validated against real-world knowledge and skill breakdowns.

7.3.2 Feedback enabled by RDF scoring

A primary advantage of the RDF scoring approach is the potential to provide feedback on specific elements of the rubric (and by design, the underlying aspect of the construct being measured), as illustrated in the following two example reports.

Figure 7-1 shows a report that displays a response from Scenario 3 where scoring has enabled the display of specific elements of the response in association with specific aspects of the rubric. This identifies specific content and points for claim and evidence, with the total score and overall evaluation described and explained. A more extended version of the report could also have included lists of the potential evidence that were not used in the response, explaining the gap between the score of 7/12 and the possible 12/12.

Figure 7-1. Example of Detailed Scenario 2 Score Report

Detailed Score Report and Feedback
HT Item Response - Claim + Evidence Score
3705 - Jane Doe 18 August 2016

tre_group	Loc	Sentence begins with…	tre_explanation	point value
claim	1	One of Harriet's most essential …	Strong claim, identifies a trait as highly important.	4
evidence	2	They all trusted her and coop …	Notes that HT was trusted by the slaves.	1
evidence	3	Throughout the journey, Harriet …	HT, at times, was tired, hungrey and afraid, just like the slaves.	1
evidence	4	She knows what she is doing …	Notes HT will do anything to help her group succeed.	1

Category	Max Points	Basis
Claim	4	Statement of one most essential trait for maximum credit; fewer points for partial response.
Evidence and Reasoning	8	Identifying attitudes, thoughts and actions that demonstrate the trait or connect it to leadership.

Raw Score Range	Final Scaled Score	Final Score Descriptor
12	4	Exceptional Achievement. Response contains all suggested elements and reflects a thorough understanding of the question and related texts.
8 - 12	3	Commendable / Adequate Achievement. Response contains many or all of the suggested elements reflecting a good understanding of the question and the related texts.
5 - 7	2	Some Achievement. Response contains an answer having some of the suggested elements, reflecting a basic understanding of the question and the related text.
1 - 4	1	Minimal / Little Evidence of Achievement. Response contains minimally relevant information reflecting at best a partial understanding of the question or of the texts.
0	0	No evidence of understanding the question or the related texts.

3705 - Moon 10H_4G1031806

Leaders all around the world and in history had great qualities in order for them to become a leader. One great leader in history is the abolitionist, Harriet Tubman. She helped free slaves for many years because they all deserved to be free. Although, without the cooperation with the slaves, Harriet Tubman wouldn't have been able to help free them. One of Harriet's most essential quality was that she was trustworthy. (1) The slaves entrusted her with their safety and is why they were able to escape slavery. They entrusted their lives to Harriet Tubman and shows that Harriet Tubman was a trustworthy person.

When Harriet gathered up a group of slaves, she promised them that they be able to get food and shelter. Even though they had very rough conditions, Harriet knew that they would be able to provide the resources they need. "She had told them about the place where they would stay, promising warmth and good food, holding these things out to the as an incentive to keep going." (The Railroad Runs to Canada). Harriet makes a promise to them in order for them to keep going and following her. The slaves believed Harriet when she promised them food and shelter. They all had to believe that they will find the resources. Harriet wasn't sure about keeping her promise, but she had to believe in order for the slaves to trust her.

As the leader, Harriet had to make sure that all of her followers trust her. Unfortunately, one slave spoke up and said that they wanted to go back. Although, Harriet wouldn't allow that to happen. She knows what could possibly happen if a slave returned. "If a runaway returned, he would turn traitor, the master and the overseer would force him to turn traitor." (The Railroad Runs to Canada). She explains to the group the possibilities of returning and they all believed her because she is the more experienced (4) out of them all. She knows what she is doing and she needs the slaves to believe in her. That is why Harriet is trusted as leader because she has more knowledge of what is going on unlike the slaves.

Throughout the journey, Harriet would get tired just like the rest of the slaves. (3) She would sleep whenever she wasn't able to walk anymore. While she sleeps, the slaves wait for her to wake up since she is the leader. "She was leading them into freedom, and so they waited until she was ready to go on." (The Railroad Runs to Canada). Her followers know that she is the only one who can guide them to freedom. The believe that she would take them to a better place so that they could all be free. They trust her with their lives so much that they respect her resting time. "A good leader "walks the talk" and in doing so earns the right to have responsibility for others." (Seven Qualities of a Good Leader). Harriet earned their respect, therefore, the slaves will wait for her to wake up. They would only allow Harriet to guide them and to tell them what to do. She's placed a huge impact on their lives and entrusts her with their freedom.

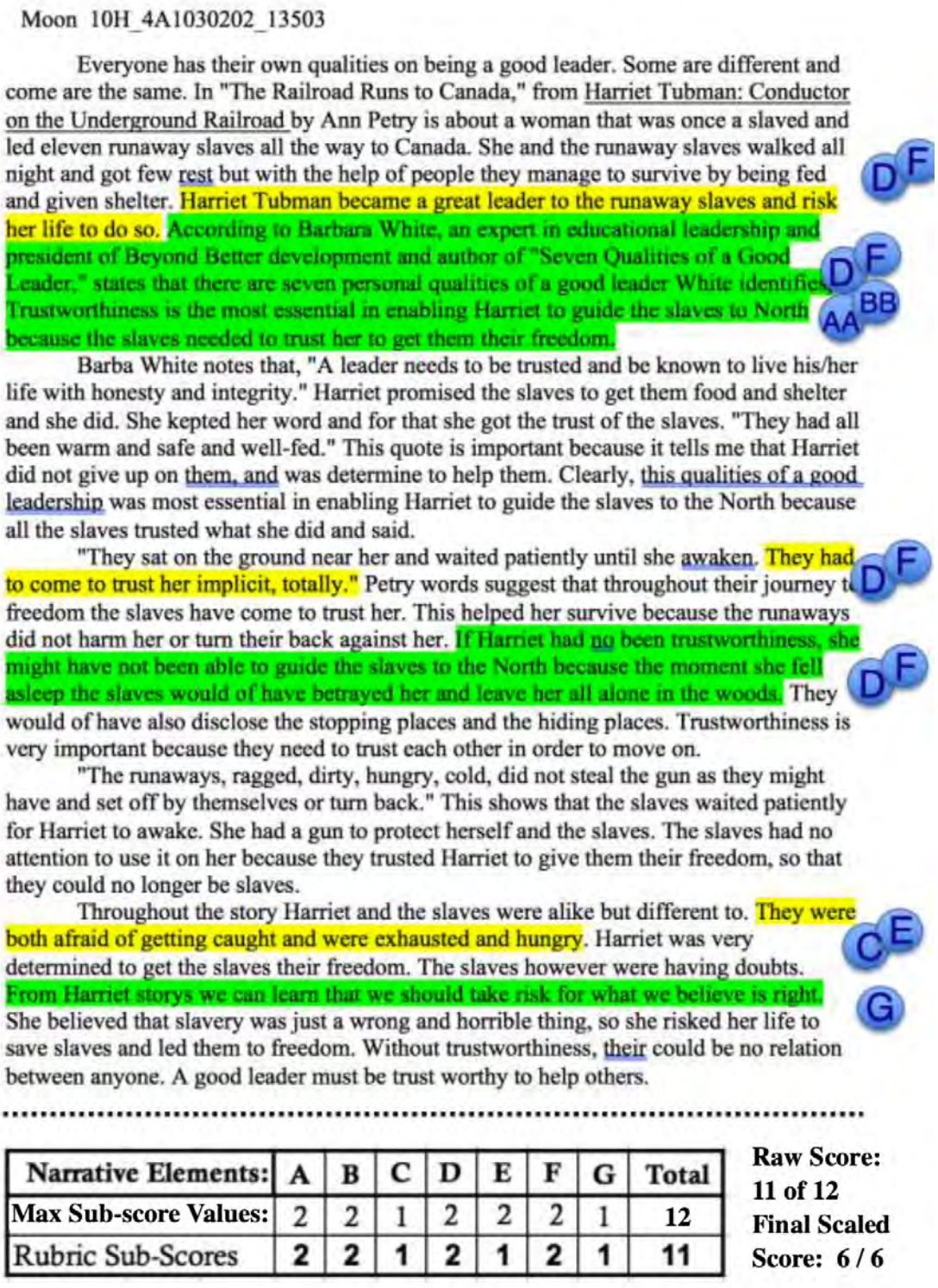
As a leader, Harriet proved how great she was by leading the slaves to freedom. (2) They all trusted her and cooperated with her in order to be free. Through her actions of freeing the slaves, Harriet shows that one of her most essential qualities is that she is trustworthy.

Scoring Results Key:

Claim score: 4/4 Evidence Score: 3/8 **Total score: 4 + 3 = 7 / 12**

The second example in Figure 7-2 shows a potential score report for a response scored in Scenario 3, where sentences tagged as representing each of the A–G narrative elements are marked for credit.

Figure 7-2. Example of Detailed Scenario 3 Score Report



Moon 10H_4A1030202_13503

Everyone has their own qualities on being a good leader. Some are different and come are the same. In "The Railroad Runs to Canada," from Harriet Tubman: Conductor on the Underground Railroad by Ann Petry is about a woman that was once a slaved and led eleven runaway slaves all the way to Canada. She and the runaway slaves walked all night and got few rest but with the help of people they manage to survive by being fed and given shelter. Harriet Tubman became a great leader to the runaway slaves and risk her life to do so. According to Barbara White, an expert in educational leadership and president of Beyond Better development and author of "Seven Qualities of a Good Leader," states that there are seven personal qualities of a good leader White identifies. Trustworthiness is the most essential in enabling Harriet to guide the slaves to North because the slaves needed to trust her to get them their freedom.

Barba White notes that, "A leader needs to be trusted and be known to live his/her life with honesty and integrity." Harriet promised the slaves to get them food and shelter and she did. She kepted her word and for that she got the trust of the slaves. "They had all been warm and safe and well-fed." This quote is important because it tells me that Harriet did not give up on them, and was determine to help them. Clearly, this qualities of a good leadership was most essential in enabling Harriet to guide the slaves to the North because all the slaves trusted what she did and said.

"They sat on the ground near her and waited patiently until she awaken. They had to come to trust her implicit, totally." Petry words suggest that throughout their journey t freedom the slaves have come to trust her. This helped her survive because the runaways did not harm her or turn their back against her. If Harriet had no been trustworthiness, she might have not been able to guide the slaves to the North because the moment she fell asleep the slaves would of have betrayed her and leave her all alone in the woods. They would of have also disclose the stopping places and the hiding places. Trustworthiness is very important because they need to trust each other in order to move on.

"The runaways, ragged, dirty, hungry, cold, did not steal the gun as they might have and set off by themselves or turn back." This shows that the slaves waited patiently for Harriet to awake. She had a gun to protect herself and the slaves. The slaves had no attention to use it on her because they trusted Harriet to give them their freedom, so that they could no longer be slaves.

Throughout the story Harriet and the slaves were alike but different to. They were both afraid of getting caught and were exhausted and hungry. Harriet was very determined to get the slaves their freedom. The slaves however were having doubts. From Harriet storys we can learn that we should take risk for what we believe is right. She believed that slavery was just a wrong and horrible thing, so she risked her life to save slaves and led them to freedom. Without trustworthiness, their could be no relation between anyone. A good leader must be trust worthy to help others.

Narrative Elements:	A	B	C	D	E	F	G	Total
Max Sub-score Values:	2	2	1	2	2	2	1	12
Rubric Sub-Scores	2	2	1	2	1	2	1	11

Raw Score: 11 of 12
Final Scaled Score: 6 / 6

This report leaves no ambiguity around why the scores were assigned and how the overall score was determined. A more complete form of the report would show an additional explanatory page where the meaning and purpose of each of the A–G elements were defined.

In both of these examples, I have illustrated how the scoring information that associated the rubric quality level definitions with specific content in the response plays a key role in explaining the score, educating the student and identifying potential gaps between performance achieved and performance that is possible.

7.3.3 Score defensibility

The same mechanism that enables the detailed score reports with feedback and transparency also provides an implicit defence of scores assigned. This element, while obvious and implicit in the other factors, is highlighted primarily to address or emphasise the advantage of RDF scoring over the challenges of defending holistic scores that have collapsed a broad range of factors into a single descriptive number. This factor is already reshaping large-scale standardised tests, where writing scores in particular are uniformly moving to multi-trait scoring (see two examples cited in Section 1.2).

7.3.4 Prerequisite deficits

For multifaceted assessment items that consider several factors among many, RDF rubrics can easily be constructed to identify a range of scorable features (e.g., writing mechanics, grammar) as distinct from others (reasoning and argumentation), simplifying the process of differentiating responses that fail to address prerequisite criteria from those that can be scored with nuance based upon the higher-level cognitive challenges posed by the item. Scoring instructions could guide scorers in marking an exam to limit scoring to prerequisite criteria if scoring falls below some threshold, allowing faster and less expensive scoring by bypassing more challenging aspects of the assessment, and the scoring, that are likely to be of limited utility for students not yet ready for such work.

7.4 Limitations

7.4.1 New rubrics may reflect a different construct

The first limitation is that the items repurposed were scored with new trial rubrics that have not themselves been validated in terms of construct representation. Nor are the items and item responses ideal for this use: The RDF rubric in Scenario 1 reflects what I believe is a solid and near-equivalent of the original holistic rubric, but this has not been validated and others might read the item differently. Such distinctions,

however, do not detract from the main focus of this study – the many benefits of response specific feedback and high inter-rater reliability that this rubric structure enables.

7.4.2 Number and variety of item types was limited

Another limitation of this study relates to the size and variety of item types and response lengths investigated. While the three scenarios included common elements in CT and AW assessment, they reflect a limited number of constructs, item types, response lengths, and students that is relatively small across many dimensions. However, by using item responses of varying lengths, and rubrics of varying levels of specificity, this study lays sufficient groundwork for the investigation of broader applications.

7.4.3 Correlational nature of much of the analyses

A further limitation related to the study design is the correlational nature of much of the analyses, which also can limit these findings in terms of explicating causal mechanisms accounting for the results. I have through careful observation done work to highlight common issues in scoring across the item types and rubrics, but a research design informed by these preliminary findings could go further by using items with externally validated construct representations for distinct subscores. In particular, quality level definitions present an ideal focus for validation work, insuring that scoring levels translate well to observable examinee abilities and knowledge.

7.4.4 Small sample size

Significant effort went into gathering data for this project, and several avenues required significant resources and yielded little results so far. The study worked with, overall, 160 item responses for Scenario 1 and 120 item responses in Scenarios 2 and 3). Four scorers were trained with the RDF rubric and three contributed scores to the detailed analysis recorded here. The analysis presented in this study reflects the first pairs of original scorers that worked on each scenario. Additional work after this study was completed resulted in additional scoring done on each data set, and I was pleased to see even higher correlations between scorers as more they gained more experience - resulting in final scaled score QWK IRR measures above 0.90. This validation notwithstanding, a limitation of the study is both the diversity of item types, number

of responses scored and number of scorers. That said, the selected scenarios reflect a range of item response lengths and questions types that are representative of actual assessments used educational settings.

7.4.5 No measure of intra-rater reliability

The study did not include actions that would allow the calculation of intra-rater reliability as well. However, the lack of this information did not detract from the central question of enabling improved feedback from associating item response elements to rubric components whilst maintaining or improving inter-rater reliability.

7.4.6 No detailed scoring reports

Finally, the quality of feedback that could be generated from the detailed scoring data collected in Scenarios 2 and 3 as scores were assigned by raters to specific sentence elements in the response was not deeply investigated during this study. The clear ability to provide specific acknowledgement of correct response characteristics is fairly intuitive and obvious. The ability to provide further feedback based on the collective set of scoring observations, including opportunities for score improvement, could yet be further explored and documented with more sophisticated score reporting software. Relatedly, more robust rubrics could have demonstrated the utility of identifying and down-scoring responses for common misconceptions, particularly useful in CT scoring embedded in subject domains such as biology and physics (see Neham and Reilly, 2007; Akmam *et al.*, 2018). In addition, data captured by the remote scoring platform created for this study is sufficient to generate reports that document both what rubric elements were satisfied, but which were not, enabling the robust feedback only illustrated with sample reports in this chapter.

7.5 Suggestions for Further Research

An examination of the results from the scoring study's three scenarios supports the conclusion that using the RDF developed in this work to build CT and AW rubrics could lead to better feedback, more transparent and defensible scoring, and a better learning experience for students given CR items for CT and AW assessment. The framework itself for describing a rubric is included in a generic form as Appendix J to this dissertation. This study suggests that this improved feedback can be created without a significant decrease in IRR, and that data generated by scoring could be

used to significantly improve feedback and understanding of the scores, as compared with holistic scoring.

Suggestions for further research to address the different shortcomings and limitations in the study, and to examine its success more closely, could include:

- **Validated Rubrics re-cast in RDF terms**. This research could easily be replicated with items whose holistic rubrics reflect the focus on CT or AW. Creating RDF versions of an existing CT rubric would more fully validate the success found with the items in this study.

- **Larger scale studies**. Trials with greater variety of item types and larger number of students, particularly if leveraging existing CT or AW assessment results, would provide an excellent approach to validating the results of this study.

- **Explore the potential for human mediated automated feedback**. As many detailed quality level rubrics, evaluated and taken together, will often inform both what is good and what is inadequate in an assessment response, it could be possible to automate, or partially automate, detailed feedback from RDF rubrics for CT. Evaluating the effectiveness of this feedback, and its success in achieving "assessment as learning" is a worthy research goal that seems within reach.

- **Intervention studies**. The notion illustrated in this study—to analyse the same responses with different rubrics—could also be used to interrogate the relative effectiveness and success of CT/AW interventions, exploring the degree to which specific rhetorical strategies taught (e.g., the narrative elements rubric) were reflected in responses, and the degree to which they correlate with the positive scores for either discrete aspects of CT or AW (e.g., strength of evidence, correctness of claim) or the overall success with the challenge.

- **Broader application**. This work appears to have broader applicability to other sorts of text scoring I have participated in over the last several years,

including such diverse areas as psychological profiling, diagnostic radiology, sales effectiveness training and compliance monitoring, where the framework's support for non-trivial algorithms to calculate subscores or combine subscores for overall performance indicators is particularly important. For example, in diagnostic radiology assessment, a "fatal miss" reading a three-dimensional image scan despite other cogent and accurate observations would outweigh any simple algebraic combination of sub-scores.

7.6 Conclusion

The research question guiding this study was:

Can a generalised and flexible RDF for scoring CT items (as compared to generic, holistic rubrics) be successfully used to define item-specific, content-centric rubrics that can guide essay graders to provide

- useful feedback to students and teachers;
- nuanced scoring that makes the exercise a learning experience;
- explicit, defensible rationales for scoring outcomes; and
- better interrater reliability?

The result of this works makes a clear case that my rubric design framework provides a rigorous, flexible, robust and generalisable mechanism to achieve these ends for CR items assessing CT and AW skills.

The rigour is enabled by structured, detailed quality level definitions for each quality level, for each evaluative criterion or sub-scores, defined for a given rubric / assessment and construct.

Flexibility was illustrated in the suitability for the schema to represent the categorical notions of "narrative elements" for content categories in Scenario 3 as easily as it represented the discrete, content-specific observations that supported the notion of the WH adapting to a new environment in Scenario 1. Flexibility was further on display

when adjustments to rubrics between development and testing phases could easily be accommodated by the same schema.

The Rubric Design Framework for CT and AW is robust by virtue of the combination of both the structure of the subscore, quality level, quality level definition hierarchies and flexibility provide in the subscale score calculation formula, the final raw score formula and the final scaling formula aspects of the RDF. While not leveraged for the items in this study, other work I have done in scoring text for such diverse areas as psychological profiling, diagnostic radiology, sales effectiveness training and compliance monitoring can address constructs where the combinations and interdependencies between subscores can benefit from more complex algorithms for score calculation.

And finally, the generalizability of the RDF structure is reflected inside the CT measurement space itself, where a construct that is sometimes measured along dozens of different aspects can well be accommodated with rubrics using the nearly limitless flexibility supported by the framework.

The secondary research question asked: Are there aspects of scoring with item-specific, content-centric rubrics that work well or that make scoring easier or more efficient?

Based on the work of this study, fast and easy scoring is possible for at least a class of challenges that use robust, content-centric, item specific quality level definitions. Such items will have straightforward and specific criteria for evidence and claims that can be expressed in an item specific and content-centric way that can lead to scoring that is described as fast and easy by scorers, and provides defensible scoring results, useful feedback to students and strong inter-rater reliability often associated with simple, holistic scoring.

This study adds to our understanding of and appreciation for item-specific rubrics, which by harmonizing a common understanding of the relative merits of differential claim and evidence citations, minimizes differing scorer judgements and enables improved feedback to students and defensible scoring outcomes, with the potential for

improved interrater reliability. In this way the study achieves the objectives for analytic rubrics stressed by Nordrum, Evans and Gustafsson (2013) for rubric-articulated feedback while addressing interrater reliability challenges for analytic CT scoring noted by Saxton, Belanger and Becker (2012).

<<<the end>>>

Appendix A: WH Item Materials

The source essay, the writing prompt, and the original holistic rubric for the HT item used in Scenario 1 are provided below.

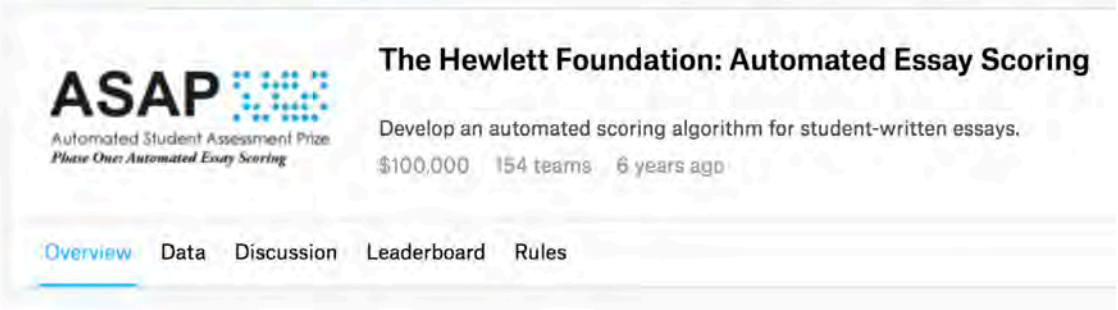
Winter Hibiscus Item Materials

Essay Set #4

Type of essay:	Source Dependent Responses
Grade level:	10
Training set size:	1,772 essays
Final evaluation set size:	589 essays
Average length of essays:	150 words
Scoring:	1st Reader Score, 2nd Reader Score, Resolved CR Score
Rubric range:	0-3
Resolved CR score range:	0-3

Retrieved from the Kaggle.com web site that hosted the ASAP contest in 2012, as described here:

https://www.kaggle.com/c/asap-aes



Data here:

https://www.kaggle.com/c/asap-aes/data

Essay Set #4 1

See https://www.kaggle.com/c/asap-aes.

A1. **WH Item Passage**

Source Essay

Winter Hibiscus by Minfong Ho

Saeng, a teenage girl, and her family have moved to the United States from Vietnam. As Saeng walks home after failing her driver's test, she sees a familiar plant. Later, she goes to a florist shop to see if the plant can be purchased.

It was like walking into another world. A hot, moist world exploding with greenery. Huge flat leaves, delicate wisps of tendrils, ferns and fronds and vines of all shades and shapes grew in seemingly random profusion.

"Over there, in the corner, the hibiscus. Is that what you mean?" The florist pointed at a leafy potted plant by the corner.

There, in a shaft of the wan afternoon sunlight, was a single blood-red blossom, its five petals splayed back to reveal a long stamen tipped with yellow pollen. Saeng felt a shock of recognition so intense, it was almost visceral.

"Saebba," Saeng whispered.

A saebba hedge, tall and lush, had surrounded their garden, its lush green leaves dotted with vermilion flowers. And sometimes after a monsoon rain, a blossom or two would have blown into the well, so that when she drew the well water, she would find a red blossom floating in the bucket.

Slowly, Saeng walked down the narrow aisle toward the hibiscus. Orchids, lanna bushes, oleanders, elephant ear begonias, and bougainvillea vines surrounded her. Plants that she had not even realized she had known but had forgotten drew her back into her childhood world.

When she got to the hibiscus, she reached out and touched a petal gently. It felt smooth and cool, with a hint of velvet toward the center—just as she had known it would feel.

And beside it was yet another old friend, a small shrub with waxy leaves and dainty flowers with purplish petals and white centers. "Madagascar periwinkle," its tag announced. How strange to see it in a pot, Saeng thought. Back home it just grew wild, jutting out from the cracks in brick walls or between tiled roofs.

And that rich, sweet scent—that was familiar, too. Saeng scanned the greenery around her and found a tall, gangly plant with exquisite little white blossoms on it. "Dok Malik," she said, savoring the feel of the word on her tongue, even as she silently noted the English name on its tag, "jasmine."

One of the blossoms had fallen off, and carefully Saeng picked it up and smelled it. She closed her eyes and breathed in, deeply. The familiar fragrance filled her lungs, and Saeng could almost feel the light strands of her grandmother's long gray hair, freshly washed, as she combed it out with the fine-toothed buffalo-horn comb. And when the sun had dried it, Saeng would help the gnarled old fingers knot the hair into a bun, then slip a dok Malik bud into it.

Saeng looked at the white bud in her hand now, small and fragile. Gently, she closed her palm around it and held it tight. That, at least, she could hold on to. But where was the fine-toothed comb? The hibiscus hedge? The well? Her gentle grandmother?

A wave of loss so deep and strong that it stung Saeng's eyes now swept over her. A blink, a channel switch, a boat ride into the night, and it was all gone. Irretrievably, irrevocably gone.

And in the warm moist shelter of the greenhouse, Saeng broke down and wept.

It was already dusk when Saeng reached home. The wind was blowing harder, tearing off the last remnants of green in the chicory weeds that were growing out of the cracks in the sidewalk. As if oblivious to the cold, her mother was still out in the vegetable garden, digging up the last of the onions with a rusty trowel. She did not see Saeng until the girl had quietly knelt down next to her.

Her smile of welcome warmed Saeng. "Ghup ma laio le? You're back?" she said cheerfully. "Goodness, it's past five. What took you so long? How did it go? Did you—?" Then she noticed the potted plant that Saeng was holding, its leaves quivering in the wind.

Mrs. Panouvong uttered a small cry of surprise and delight. "Dok faeng-noi!" she said. "Where did you get it?"

"I bought it," Saeng answered, dreading her mother's next question.

"How much?"

For answer Saeng handed her mother some coins.

"That's all?" Mrs. Panouvong said, appalled, "Oh, but I forgot! You and the

Lambert boy ate Bee-Maags . . . ."

"No, we didn't, Mother," Saeng said.

"Then what else—?"

"Nothing else. I paid over nineteen dollars for it."

"You what?" Her mother stared at her incredulously. "But how could you? All the seeds for this vegetable garden didn't cost that much! You know how much we—" She paused, as she noticed the tearstains on her daughter's cheeks and her puffy eyes.

"What happened?" she asked, more gently.

"I—I failed the test," Saeng said.

For a long moment Mrs. Panouvong said nothing. Saeng did not dare look her mother in the eye. Instead, she stared at the hibiscus plant and nervously tore off a leaf, shredding it to bits.

Her mother reached out and brushed the fragments of green off Saeng's hands. "It's a beautiful plant, this dok faeng-noi," she finally said. "I'm glad you got it."

"It's—it's not a real one," Saeng mumbled.

"I mean, not like the kind we had at—at—" She found that she was still too shaky to say the words at home, lest she burst into tears again. "Not like the kind we had before," she said.

"I know," her mother said quietly. "I've seen this kind blooming along the lake. Its flowers aren't as pretty, but it's strong enough to make it through the cold months here, this winter hibiscus. That's what matters."

She tipped the pot and deftly eased the ball of soil out, balancing the rest of the plant in her other hand. "Look how root-bound it is, poor thing," she said. "Let's plant it, right now."

She went over to the corner of the vegetable patch and started to dig a hole in the ground. The soil was cold and hard, and she had trouble thrusting the shovel into it. Wisps of her gray hair trailed out in the breeze, and her slight frown deepened the wrinkles around her eyes. There was a frail, wiry beauty to her that touched Saeng deeply.

"Here, let me help, Mother," she offered, getting up and taking the shovel away from her.

Mrs. Panouvong made no resistance. "I'll bring in the hot peppers and bitter melons, then, and start dinner. How would you like an omelet with slices of the bitter melon?"

"I'd love it," Saeng said.

Left alone in the garden, Saeng dug out a hole and carefully lowered the "winter hibiscus" into it. She could hear the sounds of cooking from the kitchen now, the beating of eggs against a bowl, the sizzle of hot oil in the pan. The pungent smell of bitter melon wafted out, and Saeng's mouth watered. It was a cultivated taste, she had discovered—none of her classmates or friends, not even Mrs. Lambert, liked it—this sharp, bitter melon that left a golden aftertaste on the tongue. But she had grown up eating it and, she admitted to herself, much preferred it to a Big Mac.

The "winter hibiscus" was in the ground now, and Saeng tamped down the soil around it. Overhead, a flock of Canada geese flew by, their faint honks clear and—yes—familiar to Saeng now. Almost reluctantly, she realized that many of the things that she had thought of as strange before had become, through the quiet repetition of season upon season, almost familiar to her now. Like the geese. She lifted her head and watched as their distinctive V was etched against the evening sky, slowly fading into the distance.

When they come back, Saeng vowed silently to herself, in the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take that test again.

A2. WH Item Prompt and Rubric

Prompt

Read the last paragraph of the story.

"When they come back, Saeng vowed silently to herself, in the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take that test again."

Write a response that explains why the author concludes the story with this paragraph. In your response, include details and examples from the story that support your ideas.

Rubric Guidelines

Score 3: The response demonstrates an understanding of the complexities of the text.

- Addresses the demands of the question
- Uses expressed and implied information from the text
- Clarifies and extends understanding beyond the literal

Score 2: The response demonstrates a partial or literal understanding of the text.

- Addresses the demands of the question, although may not develop all parts equally
- Uses some expressed or implied information from the text to demonstrate understanding
- May not fully connect the support to a conclusion or assertion made about the text(s)

Score 1: The response shows evidence of a minimal understanding of the text.

- May show evidence that some meaning has been derived from the text
- May indicate a misreading of the text or the question
- May lack information or explanation to support an understanding of the text in relation to the question

Score 0: The response is completely irrelevant or incorrect, or there is no response.

Adjudication Rules

- If Reader-1 Score and Reader-2 Score are exact or adjacent, adjudication by a third reader is not required.
- If Reader-1 Score and Reader-2 Score are not adjacent or exact, then adjudication by a third reader is required.

Appendix B: WH Initial RDF Rubric and Scoring Instructions

This appendix contains instructions for scoring the WH item with the original (Phase 1) RDF rubric.

WH RDF Scoring Instructions

Scoring instructions for graders using initial Phase 1 RDF rubric are listed below. The final section of notes was added after initial scorer training and after scoring the phase one scoring activity.

Propositions/conclusions. A set of four possible claim scores is defined below, with a decreasing point value of 16/12/8/4 and zero.

1. The final paragraph signals the significance and teenager's recognition of the overarching analogy the story communicates: that of the adaptation of the winter hibiscus to its environment and the struggle required by the immigrant teenage girl to adapt to her new environment. Adaptation and survival, 'that is what matters', her mother had said. **[Articulates central elements of the underlying analogy between adaptation of the hibiscus to a new land and the adaptation of immigrants to their new land.]**

2. The final paragraph reinforces the notion that adaptation is matter of both struggle and accommodation, and the adapter is changed in the process—becoming stronger and yet different. **[Articulates important elements of the underlying central analogy: adaptation, for the winter hibiscus or the immigrants, requires work, change, accommodation, and growth.]**

3. The final paragraph signals Saeng's determination to adapt and succeed in her new life, irrespective of whether she passes the test. She must try, and adapt and survive, because 'that is what matters'. **[Articulates some elements of the underlying central analogy of adaptation for both the hibiscus and the immigrants: work, growth or determination, and the struggle to survive being paramount.]**

4. Life is about change/growth and how to respond to change. A life is made up of the series of choices persons must make as they **grow**: what to hold on to, what to treasure, and what to value—and **how to adapt**. **[Articulates one or more minimal elements of the underlying central analogy of adaptation for both the hibiscus and the immigrants: work, growth, change or determination, and the struggle to survive.]**

For explicit recognition of the analogy between adaptation of Saeng / the daughter / the narrator / the young woman / the girl to her new environment and how the winter hibiscus has adapted to its environment, award a full score. 16 points.

Responses that recognise the need to adapt / change / respond to change (like the winter hibiscus) earn 12 points.

Responses that recognise the signal of **determination** and the need to **adapt and survive** (because 'that's what matters') get 8 points.

Responses that reference growth, determination (not linked to adapting or accommodating change), or responding to change get 4 points. Recognising half the analogy—either Saeng's need to adapt or the hibiscus adaptation—also earns 5 points.

Max 16 points. Note that this scoring approach is implemented by identifying successively less specific or more generic descriptive target expressions. When scoring, the scorer should assign the best (highest point value) among multiple qualifying claims to give this portion of the score the proper value. Similarly, during examinations of evidence found in the response, there may be multiple expressions or ways to cite a single bit of evidence (e.g., 'the hibiscus is different here'), but each evidence point should be counted only once even when repeated.

Reasoning and evidence. The support evidence to be credited by the scorer can be as follows (max one point for each of six kinds of evidence numbered 1 to 6):

1. The winter hibiscus in the new place is different from the hibiscus back home.

a. 'It's not a real one. Not like the kind we had before'.

2. Winter hibiscus not as pretty.

a. Winter hibiscus is different—not as pretty, flower less beautiful than the hibiscus they knew before.

3. Winter hibiscus is strong enough to survive the cold/winter.

a. Winter hibiscus is different—stronger/more tolerant of cold/winter than the hibiscus they knew before.

4. Saeng's mother has begun to adapt to the new environment.

a. Acclimation to the cold; persevere to provide continuity for her child

5. Saeng has begun to adapt to the new environment; Saeng recognises survival requires determination and work, even change.

a. Her mother had said survival is 'what matters'.

b. Determination to succeed, do what is necessary in the new place.

6. The winter hibiscus is in some ways the same as the hibiscus back home.

a. Petals, blossoms, stamen colour/texture as before.

b. Examining the flower met expectations (feel: cool and smooth), etc.

c. Hence it has adapted/changed to accommodate the circumstances.

Scoring formula. The claim subscore should reflect the 16/12/8/4 or 0 score assigned. [Note this changed in Phase 2, collapsing the 12 and 8 scores to a single 10-point intermediate claim score.] The evidence subscore is a simple 0 to 6 points depending on which of the six categories of evidence are found in the response. The raw overall score is the sum of the two subscores, so such scores will range from [0 to 16] (for

claims) + [0 to 6] (for evidence), or overall, from 0 to 22. These scores are then cast back into the original 0 to 3-point scale, as Table B1 shows.

Table B1. Calculation of Final Score

Raw score range	Final score	Final score descriptor
13–22	3	Strong evidence of recognising and understanding the central underlying analogy of the text.
8–12	2	Some evidence of recognising and understanding the central underlying analogy of the text.
1–7	1	Minimal evidence of recognising or understanding the central underlying analogy of the text.
0	0	No evidence of recognition or understanding the central underlying analogy of the text.

Some additional notes reflecting Phase 2 scoring. For the central claim, the proposition/claim should be scored as 4, 10, or 16 in Phase 2, as the minimum 4 point score of a single bit of the underlying analogy and the full 16 points for the full analogy were consistently recognised by both scorers in Phase 1. During Phase 2, responses that recognise a significant subset of the underlying analogy—that either the immigrants or the flower had adapted, for example—will be awarded 10 points. A bit of further clarification was provided as follows.

1. A top score of 16 should be used for responses that note that hibiscus has adapted to the new environment, recognise that the immigrant(s) (Saeng alone is fine) must also adapt, and explicitly recognise this as an analogy.

2. The intermediate score of 12 (in Phase 2, 10) points can reflect responses that at least identify some sort of analogy between the immigrant(s) and the winter hibiscus, but might only remark on their common struggle, common fate, or the fact that both have to deal with cold, without explicitly recognising the big picture.

3. The intermediate score of 8 points (in Phase 2, 10) for responses that recognise an analogy between the immigrants and winter hibiscus, but it can be implicit/unstated, so long as more than one attribute of each is mentioned. For example, a response might note that the hibiscus adapts to cold but is not as pretty, or that Saeng must adapt, or grow, or overcome/struggle in her new place. Or by

identifying two attributes of WH, one of Saeng, and contains an implicit comparison. [Added in Phase 2: Any response that explicitly communicates a comparison between Saeng to the hibiscus, or her mother to the WH, also elevates claim subscore of the response to this intermediate tier.]

4. A minimally relevant answer worth 4 points only mentions growth, struggle/work, or determination as required to overcome change for either Saeng or the winter hibiscus.

Appendix C: HT Original Item Materials

C1. Pathway Project Directions - HT

"The Railroad Runs to Canada"

Pathway Project Reading and Writing Assessment

(four pages)

Directions:

You will have two class periods for this reading and writing assessment. During the first period, you will read both an excerpt from the biography of Harriet Tubman and an informational text, "Seven Qualities of a Good Leader" by educational leadership expert Barbara White. Then you will respond to several questions and engage in activities that will help you think about what you have read in preparation for writing your essay. These notes will be collected to help you and your teacher understand how well you are reading.

During the second period, you should first skim the texts to refresh your memory. Then, look over your preliminary ideas that you recorded in this packet and plan and write your essay. Allow time to review and proofread your essay and make any revisions or corrections you wish. Your essay will be evaluated both for your reading ability and your writing ability.

Please write your name here. Your notes will be returned to you tomorrow and you can refer to them when you write your essay.

Reread the texts silently. There is room on the pages for you to mark up the text as you read. You may make notations such as these:

- Make notes about any details that stand out
- Write questions you have about the use of certain words or phrases
- Make notes about anything that is similar to your own experiences
- Comment on parts that you think are especially interesting

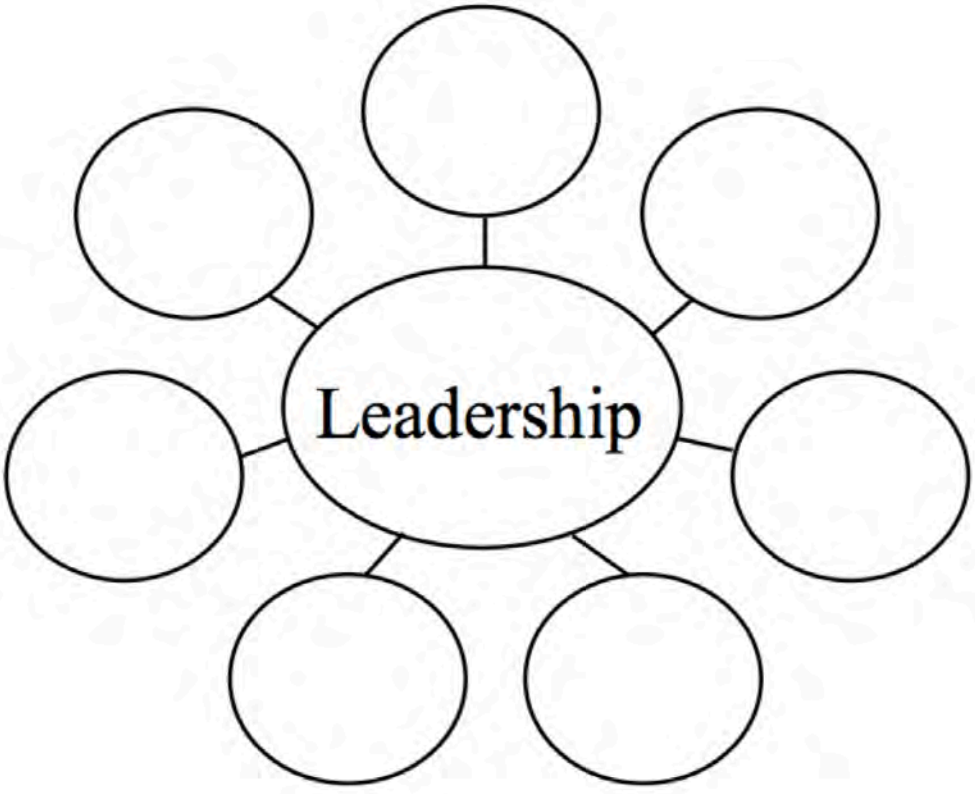
C2. HT Item Instructions

"The Railroad Runs to Canada"

Examining Leadership in *The Railroad Runs to Canada*

You have just read an excerpt from the biography of Harriet Tubman and an article called "Seven Qualities of a Good Leader." The following activities will help you apply what you've learned about leadership to the character of Harriet Tubman in "The Railroad Runs to Canada.".

1. Use White's article "Seven Qualities of a Good Leader" to list seven qualities of leadership in the cluster below. Then star the characteristic that you believe was most essential in enabling Tubman to survive.



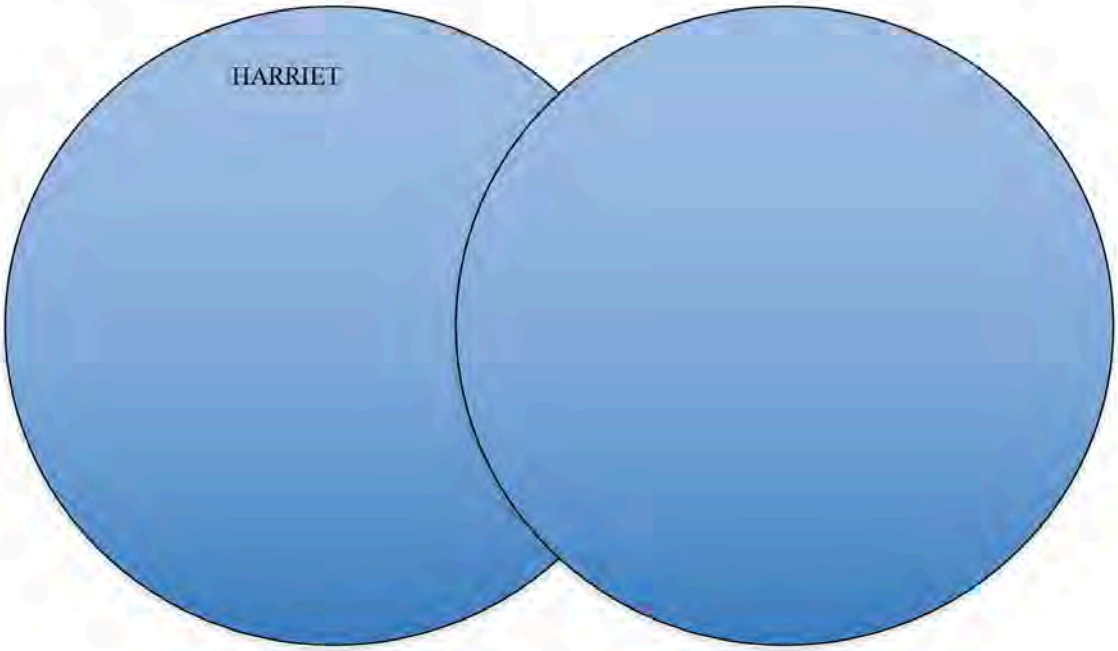
2

2. In the chart below, define the characteristic of leadership you believe was most essential in enabling Harriet to survive. Provide evidence from "The Railroad Runs to Canada" (examples and direct quotes) that illustrate how Harriet demonstrates that characteristic. Then explain why you feel this trait of leadership was so critical to Harriet's survival.

CHARACTERISTIC OF LEADERSHIP MOST ESSENTIAL TO HARRIET'S SURVIVAL:

Definition of this characteristic:	
1) EVIDENCE FROM "THE RAILROAD RUNS TO CANADA":	1) HOW THIS HELPED HARRIET SURVIVE:
2) EVIDENCE FROM "THE RAILROAD RUNS TO CANADA":	2) HOW THIS HELPED HARRIET SURVIVE:

3. Use the graphic organizer below to compare and contrast Harriet's response to the life-threatening situation with the responses of her followers. In what ways was Harriet similar to her follower and in what ways was she different? How did these differences contribute to her survival?



4. WRITING YOUR CLAIM:
Which characteristic of leadership was most essential in enabling Harriet to survive and emerge as a leader?

C3. HT Item Passage

Student Name:_____

NOTES/REACTIONS

The Railroad Runs to Canada

from *Harriet Tubman: Conductor on the*

[Harriet Tubman: Conductor on the Underground Railroad?]

Underground Railroad by Ann Petry

Harriet Tubman 1820-1913

Harriet Tubman was born into slavery in Maryland in 1820 and successfully escaped in 1849. Rather than remaining in the safety of the North, Harriet made it her mission to return to the South to rescue family members and others living in slavery via the Underground Railroad, an elaborate network of safe houses organized to help slaves reach free states. This excerpt documents what happened in 1851, when Harriet led eleven runaway slaves all the way to Canada.

Along the Eastern Shore of Maryland, in Dorchester County, in Caroline County, the masters kept hearing whispers about the man named Moses, who was running off slaves. At first they did not believe in his existence. The stories about him were fantastic, unbelievable. Yet they watched for him. They offered rewards for his capture.

They never saw him. Now and then they heard whispered rumors to the effect that he was in the neighborhood. The woods were searched. The roads were watched. There was never anything to indicate his whereabouts. But a few days afterward, a goodly number of slaves would be gone from the plantation. Neither the master nor the overseer had heard or seen anything unusual in the quarter. Sometimes one or the other would vaguely remember having heard a whippoorwill call somewhere in the woods, close by, late at night. Though it was the wrong season for whippoorwills… There was never anything more than that to suggest that all was not well in the quarter. Yet, when morning came, they invariably discovered that a group of the finest slaves had taken to their heels.

Unfortunately, the discovery was almost always made on a Sunday. Thus a whole day was lost before the machinery of pursuit could be set in motion. The posters offering rewards for the fugitives could not be printed until Monday. The men who made a living hunting for runaway slaves were out of reach, off in the woods with their dogs and their guns, in pursuit of four-footed game, or they were in camp meetings saying their prayers with their wives and families beside them.

Harriet Tubman could have told them that there was far more involved in this matter of running off slaves than signaling the would-be runaways by imitating the call of a whippoorwill, or a hoot owl, far more involved than a matter of waiting for a clear night when the North Star was visible.

In December 1851, when she started out with the band of fugitives that she planned to take to Canada, she had been in the vicinity of the plantation for days, planning the trip, carefully selecting the slaves that she would take with her. She had announced her arrival in the quarter by singing the forbidden spiritual—"Go down, Moses, 'way down to Egypt Land"—singing it softly outside the door of a slave cabin, late at night. The husky voice was beautiful even when it was barely more than a murmur borne on the wind.

Student Name:_____

NOTES/REACTIONS

Note: *Underlined words are defined at the back of this excerpt.*

Once she had made her presence known, word of her coming spread from cabin to cabin. The slaves whispered to each other, ear to mouth, mouth to ear, "Moses is here." "Moses has come." "Get ready. Moses is back again." The ones who had agreed to go North with her put ashcake and salt herring in an old bandanna, hastily tied it into a bundle, and then waited patiently for the signal that meant it was time to start.

There were eleven in this party, including one of her brothers and his wife. It was the largest group that she had ever conducted, but she was determined that more and more slaves should know what freedom was like. She had to take them all the way to Canada. The Fugitive Slave Law was no longer a great many incomprehensible words written down on the country's lawbooks. The new law had become a reality. It was Thomas Sims, a boy, picked up on the streets of Boston at night and shipped back to Georgia. It was Jerry and Shadrach, arrested and jailed with no warning.

She had never been in Canada. The route beyond Philadelphia was strange to her. But she could not let the runaways who accompanied her know this. As they walked along, she told them stories of her own first flight; she kept painting vivid word pictures of what it would be like to be free.

But there were so many of them this time. She knew moments of doubt, when she was half afraid and kept looking back over her shoulder, imagining that she heard the sound of pursuit. They would certainly be pursued. Eleven of them. Eleven thousand dollars' worth of flesh and bone and muscle that belonged to Maryland planters. If they were caught, the eleven runaways would be whipped and sold South, but she—she would probably be hanged.

They tried to sleep during the day but they never could wholly relax into sleep. She could tell by the positions they assumed, by their restless movements. And they walked at night. Their progress was slow. It took them three nights of walking to reach the first stop. She had told them about the place where they would stay, promising warmth and good food, holding these things out to them as an incentive to keep going. When she knocked on the door of a farmhouse, a place where she and her parties of runaways had always been welcome, always been given shelter and plenty to eat, there was no answer. She knocked again, softly. A voice from within said, "Who is it?" There was fear in the voice. She knew instantly from the sound of the voice that there was something wrong. She said, "A friend with friends," the password on the Underground Railroad. The door opened, slowly. The man who stood in the doorway looked at her coldly, looked with unconcealed astonishment and fear at the eleven disheveled runaways who were standing near her. Then he shouted, "Too many, too many. It's not safe. My place was searched last week. It's not safe!" and slammed the door in her face.

She turned away from the house, frowning. She had promised her passengers food and rest and warmth, and instead of that, there would be hunger and cold and more walking over the frozen ground. Somehow she would have to instill courage into these eleven people, most of them strangers, would have to feed them on hope and bright dreams of freedom instead of the fried pork and corn bread and milk she had promised them.

They stumbled along behind her, half dead for sleep, and she urged them on, though she was as tired and as discouraged as they were. She had never been in Canada, but she kept painting wondrous word pictures of what it would be like. She managed to dispel their fear of pursuit so that they would not become hysterical, panic-stricken. Then she had to bring some of the fear back, so that they would stay awake and keep walking though they drooped with sleep. Yet, during the day, when they lay down deep in a thicket, they never really slept, because if a twig snapped or the wind sighed in the

Student Name:_____

NOTES/REACTIONS

branches of a pine tree, they jumped to their feet, afraid of their own shadows, shivering and shaking. It was very cold, but they dared not make fires because someone would see the smoke and wonder about it…

That night they reached the next stop—a farm that belonged to a German. She made the runaways take shelter behind trees at the edge of the fields before she knocked at the door. She hesitated before she approached the door, thinking, suppose that he too should refuse shelter, suppose——*Then she thought, Lord, I'm going to hold steady on to You and You've got to see me through*—and knocked softly.

She heard the familiar guttural voice say, "Who's there?"

She answered quickly, "A friend with friends."

He opened the door and greeted her warmly. "How many this time?" he asked.

"Eleven," she said and waited, doubting, wondering.

He said, "Good. Bring them in."

He and his wife fed them in the lamp-lit kitchen, their faces glowing as they offered food and more food, urging them to eat, saying there was plenty for everybody, have more milk, have more bread, have more meat.

They spent the night in the warm kitchen. They really slept, all that night and until dusk the next day. When they left, it was with reluctance. They had all been warm and safe and well-fed. It was hard to exchange the security offered by that clean, warm kitchen for the darkness and the cold of a December night.

"Go On or Die"

Harriet had found it hard to leave the warmth and friendliness, too. But she urged them on. For a while, as they walked, they seemed to carry in them a measure of contentment; some of the serenity and the cleanliness of that big, warm kitchen lingered on inside them. But as they walked farther and farther away from the warmth and the light, the cold and the darkness entered into them. They fell silent, sullen, suspicious. She waited for the moment when some one of them would turn mutinous…She told them about Frederick Douglass, the most famous of the escaped slaves, of his eloquence, of his magnificent appearance. Then she told them of her own first, vain effort at running away, evoking the memory of that miserable life she had led as a child, reliving it for a moment in the telling. But they had been tired too long, hungry too long, afraid too long, footsore too long. One of them suddenly cried out in despair, "Let me go back. It is better to be a slave than to suffer like this in order to be free."

She carried a gun with her on these trips. She had never used it—except as a threat. Now, as she aimed it, she experienced a feeling of guilt, remembering that time, years ago, when she had prayed for the death of Edward Brodas, the Master, and then, not too long afterward, had heard that great wailing cry that came from the throats of the field hands, and knew from the sound that the Master was dead.

One of the runaways said again, "Let me go back. Let me go back," and stood still, and then turned around and said, over his shoulder, "I am going back."

Student Name:_____

NOTES/REACTIONS

She lifted the gun, aimed it at the despairing slave. She said, "Go on with us or die." The husky, low-pitched voice was grim.

He hesitated for a moment and then he joined the others. They started walking again. She tried to explain to them why none of them could go back to the plantation. If a runaway returned, he would turn traitor; the master and the overseer would force him to turn traitor. The returned slave would disclose the stopping places, the hiding places, the corn stacks they had used with the full knowledge of the owner of the farm, the name of the German farmer who had fed them and sheltered them. These people who had risked their own security to help runaways would be ruined, fined, imprisoned.

She said, "We got to go free or die. And freedom's not bought with dust…"

She gave the impression of being a short, muscular, indomitable woman who could never be defeated. Yet at any moment she was liable to be seized by one of those curious fits of sleep, which might last for a few minutes or for hours. Even on this trip, she suddenly fell asleep in the woods. The runaways, ragged, dirty, hungry, cold, did not steal the gun as they might have and set off by themselves or turn back. They sat on the ground near her and waited patiently until she awakened. They had come to trust her implicitly, totally. They, too, had come to believe her repeated statement, "We got to go free or die." She was leading them into freedom, and so they waited until she was ready to go on.

By slow stages they reached Philadelphia… Harriet felt safer now, though there were danger spots ahead. But the biggest part of her job was over. As they went farther and farther north, it grew colder; she was aware of the wind on the Jersey ferry and aware of the cold damp in New York. From New York they went on to Syracuse, where the temperature was even lower.

Late in December 1851, Harriet arrived in St. Catharines, Canada West (now Ontario), with the eleven fugitives. It had taken almost a month to complete this journey… In spite of the severe cold, the hard work, she came to love St. Catharines and the other towns and cities in Canada where black men lived. She discovered that freedom meant more than the right to change jobs at will, more than the right to keep the money that one earned. It was the right to vote and to sit on juries. It was the right to be elected to office. In Canada there were black men who were county officials and members of school boards. St. Catharines had a large colony of ex-slaves, and they owned their own homes, kept them neat and clean in good repair. They lived in whatever part of town they chose and sent their children to the schools.

She continued to live in this fashion, spending the winter in Canada and the spring and summer working in Cape May, New Jersey, or in Philadelphia. She made two trips a year into slave territory, one in the fall and another in the spring. She now had a definite, crystallized purpose, and in carrying it out, her life fell into a pattern which remained unchanged for the next six years.

Harriet Tubman remained active during the Civil War, working for the Union Army as an armed scout and spy. She was the first woman to lead an armed expedition in the war and helped to liberate more than 700 slaves. After the war, she helped to lead the fight for the abolition of slavery.

Student Name:_____

NOTES/REACTIONS

Vocabulary and Notes

- masters—n., slave owners
- quarter—n., the area where the slaves lived
- taken to their heels—idiomatic expression for run away
- "Go down, Moses, 'way down to Egypt Land": a line from a well-known African-American folk song about Moses leading the enslaved Israelites out of Egypt.
- borne—adj., carried
- Fugitive Slave Law: a law passed in 1850, allowing slave owners to recover escaped slaves even if they had reached free states.
- incomprehensible; adj., impossible to understand
- disheveled—adj., messy or untidy
- instill—v., to supply gradually
- dispel—v, to drive away
- indomitable—adj., unable to be conquered
- mutinous, — adj., rebellious; a disposed to revolt against another
- eloquence – n., language that is powerful, moving, or graceful
- evoking – v., to call up or produce (such as memories)
- curious fits of sleep: mysterious spells of dizziness or unconsciousness experienced by Harriet Tubman because she was beaten and hit with a heavy metal object by a slave master
- crystallized, v., to cause something (such as an idea or belief) to become fully formed

C4. HT Ancillary Passage

Seven Qualities of a Good Leader
By Barbara White

How often have you heard the comment, "He or she is a born leader?" There are certain characteristics found in some people that seem to naturally put them in a position where they're looked up to as a leader. Whether in fact a person is born a leader or develops skills and abilities to become a leader is open for debate. However, there are some clear characteristics that are found in good leaders. Let us explore them further.

Seven Personal Qualities Found In A Good Leader:

1. **Trustworthiness**: True authority is born from respect for the good character and trustworthiness of the person who leads. A leader needs to be trusted and be known to live his/her life with honesty and integrity. A good leader "walks the talk" and in doing so earns the right to have responsibility for others.

2. **Dedication:** A good leader is enthusiastic about his/her work or cause and also about his/her role as leader. People will respond more openly to a person of passion and dedication. Leaders need to be able to be a source of inspiration, and be a motivator towards the required action or cause.

3. **Confidence:** A good leader is confident. In order to lead and set direction, a leader needs to appear confident as a person and in the leadership role. Such a person inspires confidence in others and draws out the trust and best efforts of the team to complete the task well.

4. **Organizational Skills:** A leader also needs to function in an orderly and purposeful manner in situations of uncertainty. People look to the leader during times of uncertainty and unfamiliarity and find reassurance and security when the leader portrays confidence and a positive demeanor.

5. **Steadfastness:** Good leaders are tolerant of ambiguity and remain calm, composed and steadfast to the main purpose. Storms, emotions, and crises come and go and a good leader takes these as part of the journey and keeps a cool head.

Note: *Underlined words are defined on the next page.*

6. **Analytical Skills:** A good leader as well as keeping the main goal in focus is able to think analytically. Not only is the goal in view, but a good leader can break it down into manageable steps and make progress towards it.

7. **Commitment:** A good leader is committed to excellence. The good leader not only maintains high standards, but also is proactive in raising the bar in order to achieve excellence in all areas.

These seven personal characteristics are foundational to good leadership. Some characteristics may be more naturally present in the personality of a leader. However, each of these characteristics can also be developed and strengthened.

Barbara White is an expert in educational leadership and President of Beyond Better Development*

Vocabulary

- integrity – n., the quality of being honest or fair
- demeanor – n., the way in which a person behaves; a distinguishing feature of a person's character
- ambiguity – n., uncertainty

C5. HT Prompt

The Railroad Runs to Canada
(single page)

Background
In her article 'Seven Qualities of a Good Leader,' Barbara White, author and expert in educational leadership, identifies seven key qualities that enable good leaders to guide, influence or direct others.

Writing Directions
You have just read an excerpt from Ann Petry's biography *Harriet Tubman: Conductor on the Underground Railroad,* describing how Harriet, an escaped slave, returned to southern plantations to rescue others' slaves and guide them to freedom.

PROMPT

Review White's article, 'Seven Qualities of a Good Leader.' **Write an essay in which you make a claim about ONE quality of leadership that was MOST ESSENTIAL in enabling Harriet to guide the slaves to the North.**

In the body of your essay:
Discuss how Harriet's key quality of leadership helped her to overcome several obstacles and why it was so important to her and the other slaves' survival.
Compare and contrast Harriet's response to this life-threatening situation with that of the slaves. What does Harriet share in common with her followers and what differences allowed her to emerge as a leader?

In your conclusion, describe a lesson we can learn from Harriet's story and her acts of courage.

REMEMBER to clearly address all parts of the writing task, support your main ideas with evidence from both reading selections, use precise and descriptive language, and proofread your paper to correct errors in the conventions of written English.

Appendix D: HT Original Rubric

Scoring Guide for 'The Railroad Runs to Canada' and 'Unbroken'

Note: Papers at all levels of achievement described below will contain some or all of the characteristics listed as criteria for each particular score.

6 Exceptional Achievement

1. Writer introduces the subject, giving **enough background** for the reader to follow the interpretation he/she offers in response to the prompt.
2. Writer presents **a thoughtful/insightful claim** about the quality of leadership that was most essential in enabling Harriet to inspire the slaves or the characteristic of resilience that was most essential in enabling Louie to survive.
3. Writer gives **specific examples of several obstacles Harriet and the slaves faced** and perceptively discusses how a **key leadership quality helped Harriet overcome these obstacles** or gives **specific example of several obstacles the men faced**, what Louie thought, **did, felt, and said** in response to the situation, and perceptively discusses **how his key trait of resilience helped him** to overcome these obstacles.
4. Writer thoughtfully ***compares* Harriet's response to this life-threatening situation with that of the slaves** (how she is like and different from them) or **Louie's response to that of Phil and Mac** (who is more like him and less like him).
5. Writer perceptively considers what **characteristics of leadership exhibited by Harriet** or **characteristics of resilience exhibited by Louie** reveal about each character's values and beliefs.
6. Writer thoughtfully analyses a **lesson readers can learn** from Harriet's acts of courage or Louie's story of survival.
7. Writer skillfully **weaves numerous references from both sources** (the nonfiction biography and the source materials on leadership or resilience) into the **essay to support his/her claim**.

8. Throughout the essay, writer carefully analyses the language the authors use to depict the dire circumstances the characters are in and how the language illustrates leadership or resilience.
9. **Writer uses especially precise and descriptive language as well as transition words.**
10. Writer interprets authoritatively using a formal tone and advances to a logical conclusion that clearly follows from and supports the argument presented.
11. Paper has few errors in the conventions of written English.

5 Commendable Achievement

1. Writer introduces the subject, giving enough background for the reader to follow the interpretation he/she offers in response to the prompt.
2. Writer presents a reasonably thoughtful claim about the quality of leadership that was most essential in enabling Harriet to inspire the slaves or the characteristic of resilience that was most essential in enabling Louie to survive.
3. Writer gives examples of obstacles Harriet and the slaves faced and thoughtfully discusses how a key leadership quality helped Harriet overcome these obstacles or gives examples of obstacles the men faced, what Louie thought, did, felt, and said in response to the situation, and thoughtfully discusses how his key trait of resilience helped him to overcome these obstacles.
4. Writer thoughtfully compares Harriet's response to this life-threatening situation with that of the slaves (how she is like and different from them) or Louie's response to that of Phil and Mac (who is more like him and less like him).
5. Writer thoughtfully considers what characteristics of leadership exhibited by Harriet or characteristics of resilience exhibited by Louie reveal about each characters' values and beliefs.
6. Writer thoughtfully analyses a lesson readers can learn from Harriet's acts of courage or Louie's story of survival.
7. Writer weaves some references from both sources (the nonfiction biography and the source materials on leadership or resilience) into the essay to support his/her claim.
8. Throughout the essay, writer analyses the language the authors use to depict the dire circumstances the characters are in and how the language illustrates leadership or resilience.
9. Writer uses some precise and descriptive language as well as transition words.

10. Writer interprets authoritatively using a formal tone and advances to a logical conclusion that clearly follows from and supports the argument presented, but the conclusion is less compelling than a 6 paper.
11. Paper has relatively few errors in the conventions of written English.

4 Adequate Achievement

1. Writer orients the reader adequately by giving at least some introductory context.
2. Writer may begin unsteadily but reaches a focus or point as the essay progresses.
3. Writer presents an adequate claim about the quality of leadership or characteristic of resilience that was most essential in enabling Harriet/Louie to overcome obstacles/survive.
4. Writer gives examples of obstacles Harriet and the slaves faced and discusses how a key leadership quality helped Harriet overcome these obstacles or gives examples of obstacles the men faced, what Louie thought, did, felt, and said in response to the situation, and discusses how his key trait of resilience helped him to overcome these obstacles.
5. Writer compares Harriet's response to this life-threatening situation with that of the slaves (how she is like and different from them) or Louie's response to that of Phil and Mac (who is more like him and less like him).
6. Writer considers what characteristics of leadership exhibited by Harriet or characteristics of resilience exhibited by Louie reveal about each character's values and beliefs.
7. Writer adequately analyses a lesson readers can learn from Harriet's acts of courage or Louie's story of survival.
8. Writer weaves a few references from both sources (the nonfiction biography and the source materials on leadership or resilience) into the essay to support his/her claim.
9. Writer uses less precise and descriptive language as well as transition words.
10. Writer interprets less authoritatively using a less formal tone and advances to a conclusion that supports the argument presented, but the conclusion is less compelling than a 5 or 6 paper.

3 Some Evidence of Achievement

1. Writer introduces the topic perfunctorily or simply dives in—answering the questions without developing a clear introduction.

2. Overall, writer's discussion of 'The Railroad Runs to Canada' or 'Unbroken' may be superficial or rely on the retelling of events and provide little in the way of analysis or commentary.
3. Writer may fail to make a claim about what quality of leadership or characteristic of resilience enabled Harriet to inspire the slaves or Louie to survive.
4. Writer may fail to give specific examples of the obstacles Harriet and the slaves or the men faced or give examples but fail to discuss or superficially discuss how the key trait of leadership or resilience helped Harriet/Louie to overcome obstacles.
5. Writer may fail to compare and contrast Harriet to the slaves or Louie to Phil and Mac.
6. Writer's conclusion may not connect the character's traits of leadership or resilience to his/her values and beliefs.
7. Writer may provide a superficial lesson learned or neglect to discuss what lesson can be learned.
8. Writer uses little to no precise and descriptive language or transition words.
9. Writer uses few, if any, references to the texts (the biography or nonfiction materials on leadership or resilience).
10. Paper has many errors in the conventions of written English, some of which may interfere with the writer's message.

2 Little Evidence of Achievement

1. Writer provides no introduction, or it is brief and unfocused.
2. Writer may simply retell the story without seeming to really understand everything that takes place.
3. Writer may fail to discuss characteristics of leadership and resilience and how they are demonstrated by Harriet or Louie.
4. Writer may fail to give examples of how Harriet or Louie use leadership or resilience to overcome obstacles.
5. Writer may not understand or fails to discuss the lesson learned in 'The Railroad Runs to Canada' or 'Unbroken'.
6. Writer talks in generalities and fails to provide references to the two source texts. Conclusion may be abrupt or missing.
7. Language is imprecise.

8. Paper has errors in the conventions of written English, many of which interfere with the author's message.

1 Minimal Evidence of Achievement

1. Context/introduction is missing, abrupt, or confusing.
2. Writer does not discuss or appear to understand what characteristics of leadership or resilience are displayed by Harriet or Louie.
3. Writer merely retells the story and does not describe what obstacles the characters faced or how they use leadership/resilience to overcome them.
4. Writer makes no attempt to consider what lesson can be learned from the biographies.
5. Writer fails to provide references to either the fictional text or nonfiction source material.
6. Writer has very poor command of how to construct an essay.
7. Paper has so many errors in the conventions of written English that the writer's meaning is obscured.

Appendix E: HT RDF C+E Rubric and Instructions

Claim + Evidence Rubric for Harriet Tubman Essay – Scenario 2 – HT Rubric C+E

Part 1 of Rubric: Identify expected elements of argumentation

This rubric will reward both claim and evidence in the Harriet Tubman item response essay's: an answer to the question of the primary leadership quality, and evidence provided in support of that response.

The students were asked to identify the leadership characteristic most responsible for Harriet's success from a set of seven leadership characteristics defined in a passage that accompanied the Underground Railroad excerpt.

Students were instructed to select a characteristic and justify the choice with evidence from the story about Harriet Tubman. The preparation materials ask the students to consider ways in which Harriet was similar to the enslaved peoples she was helping to find freedom, as well as differences between Harriet and her followers. Students were also asked to consider how Harriet and the followers were the same or different in their response to life-threatening situations. Finally, students were asked if there was a lesson that could be taken from the Underground Railroad story about leadership.

Scoring each response will entail associating any qualifying material in the text with either of these two types of target response elements described below. Grading and feedback will require specific parts of the text to address specific rubric elements described below.

Claim or proposition that answers the "what is the most important leadership characteristic" aspect of the prompt is part of the target response element set that will be used to grade this item.

The first part of this rubric is concerned with the student's choice of a "most important" leadership characteristic. An explicit, clear choice of a single characteristic that corresponds directly to one of the seven characteristics defined in the leadership passage that is part of the Harriet Tubman writing exercise is the expected answer. For the purposes of this exercise, any of the seven are acceptable choices. Not making a choice, not making a clear choice or selecting more than one choice are ways in which less than full credit could be received for this part of the response. The specific "best response" is not as important here as the form of argumentation being taught, and so there are no better or worse choices for the leadership characteristic, just better and worse ways of articulating a choice.

Evidence, or reasoning that connects evidence to the claim, is also expected in the response. This part of the rubric is concerned with the student's presentation of evidence to support their claim of the "most important" leadership characteristic.

1 / 5

Claim + Evidence Rubric for Harriet Tubman Essay – Scenario 2 – HT Rubric C+E

Identification of Harriet's actions and choices that manifest, support, reflect or are related to the chosen leadership characteristic, either implicitly or as asserted by the response, will be accepted as a claim of evidence. Evidence that differentiates Harriet from her followers, or contrasts differences with similarities, in service to supporting their claim are also credited. The quality of the evidence cited for a specific leadership trait is not the focus of this learning exercise, so each bit of evidence will have an equal score of one point, but points will not be awarded for citing the same evidence more than once.

The claim is scored on a 1 to 4 scale, based on one or more elements of the response.

For part one of this rubric, statements that contribute to the articulation of the primary claim or proposition that satisfies the requirement for a "most important" leadership characteristic will be given between 1 to 4 points for the strength of the claim -- in terms of its clarity and articulation.

A simple statement that "X is an important leadership characteristic" without the specificity of it being a "most important" characteristics, or the choice of two or more characteristics, or choosing a characteristic related to but no one as defined in the leadership passage would be examples of responses that warrant a lower score. Zero points should be award if no leadership characteristic is singled out.

Multiple assertions of the same claim, without clear incremental elaboration or explanation, should not provide additional points for this aspect of a response. Each sentence (or part of a sentence) that satisfies some aspect of the requirement for a claim of "most important leadership characteristic" shall receive a score of either 1 to 4 points. If portions of text span more than one sentence represent a desired response element, the point can be assigned to any sentence that is part of the relevant text.

[Note: As a simplifying assumption for our proof-of-concept implementation, feedback and scoring support will refer to an underlying sentence, and not a specific set of words. If a single key idea spans multiple sentences, the scorer should associate the sentence containing the most salient part of the text with the rubric element.]

The evidence is scored piece by piece, with each element containing evidence scored as a 1 or 2.

For the evidence part of the rubric, most citations of evidence, or argumentation connecting evidence to the claim, will be given a single score point. In cases where a single sentence combines to related observations, a score of 2 may be warranted. If two points are made across two sentences, score each sentence with one point.

Claim + Evidence Rubric for Harriet Tubman Essay - Scenario 2 - HT Rubric C+E

Part 2 of Rubric: Sum scoring micro-decisions to produce raw score total

The response elements found that are awarded points for "claim" or "evidence" will be summed within a cap for that category, and the result added together for the

total raw score. The maximum points that can be awarded for either the claim or the collective body of evidence and reasoning in a response is limited as noted in the table below. Therefore, the scoring process will sum points assigned to each category (Claim and Evidence) within the category, and apply the category maximum if necessary, before adding the resultant category scores together for the total raw score.

Maximum points per claim and evidence response element category:

Category	Claim	Evidence (and reasoning)
Max Points	4	8

Individual response elements (sentences) that satisfy rubric elements will therefor receive 1 or 2 points for Evidence elements, and between 1 and 4 points for Claim elements. For the claim, no more than 4 points will count toward the raw score, and for bits of evidence, no more than 8 points will be included in the overall raw score. [Score reports and feedback will note the full range of claim and evidence included in a response, while also showing the scoring calculation elements.]

Note that in addition to a maximum points-per-category, duplicate citations of the same evidence or claim, should not be counted. That is, each point awarded should be for a semantically distinct contribution.

To complete part 2, sum (and cap if necessary) the claim and evidence element scores and compute the total raw score which will range between zero and a maximum of 12.

The figure below shows the scoring for a single response, and that an abundance of evidence results in a "capped" overall raw score total.

See example below for Capped Raw Score for HT Item using the C+E Rubric:

Claim + Evidence Rubric for Harriet Tubman Essay - Scenario 2 - HT Rubric C+E

					max 4	max 8	max 12		
seqn	irsen_id	itemresp_id	irsen _seqn	irpara _seqn	irs_ce _cpts	irs_ce _epts	irs_tot _cepts	irs_ce _name	irsen_text
1	19365	3544	1	1	0	0	0		"Go on with u
2	19366	3544	2	1	0	1	1		Even after suc
3	19367	3544	3	1	0	0	0		Tubman's bra
4	19368	3544	4	1	4	0	6	organ	Tubman's mo
5	19369	3544	1	2	0	0	0		Harriet Tubm
6	19370	3544	2	2	0	0	0		For instance v
7	19371	3544	3	2	0	1	1		It is better to
8	19372	3544	4	2	0	1	1		She grimly sai
9	19373	3544	5	2	0	0	0		Tubman was t
10	19374	3544	6	2	0	1	1		She didn't bot
11	19375	3544	7	2	0	0	0		In the midst o
12	19376	3544	8	2	0	1	1		However, Tub
13	19377	3544	9	2	0	1	1		Tubman, like
14	19378	3544	1	3	0	0	0		Another exam
15	19379	3544	2	3	0	1	1		And freedom'
16	19380	3544	3	3	0	0	0		Those words a
17	19381	3544	4	3	0	0	0		As one of her
18	19382	3544	5	3	0	0	0		"She tried to a
19	19383	3544	6	3	0	1	1		If a runaway r
20	19384	3544	7	3	0	1	1		These differer
21	19385	3544	1	4	0	0	0		"A leader with
22	19386	3544	2	4	0	1	1		People look to
23	19387	3544	3	4	0	0	0		Tubman's org
24	19388	3544	4	4	0	0	0		A lesson that
			raw score		4	10	16		
			final score		4	8	12		
					max 4	max 8	max 12		
					claim	evidence	total CE		

Note that in the scoring figure above a total of 10 evidence points were earned by essay 3544, (which had 24 sentences) but only the max 8 evidence score points were counted toward the raw score of 12 (which was a maximum score).

Part 3 of Rubric: Determine final score

The final score will be assigned to a value equal to one half the total raw score, rounding up as shown in the table below.

Claim + Evidence Rubric for Harriet Tubman Essay - Scenario 2 - HT Rubric C+E

Final Raw Score	**Scaled Final Score**	**Final Score Descriptor**
11-12	6	Exceptional Achievement. Response contains all or nearly all of the elements suggested in the assignment, reflecting a clear understanding of the question and the related texts.
9-10	5	Commendable Achievement. Response contains many of the suggested elements reflecting a good understanding of the question and the related texts.
8 - 7	4	Adequate Achievement. Response contains an answer having some of the suggested elements, reflecting a basic understanding of the question and the related text.
5 - 6	3	Some Evidence of Achievement. Response contains some information reflecting a basic understanding of the question or the related texts.
3 - 4	2	Little Evidence of Achievement. Response contains some information reflecting a partial understanding of the question or the related texts.
1 - 2	1	Minimal or no Evidence of Achievement. Response contains minimally relevant information reflecting at best a partial understanding of the question or of the texts.
0	0	Non-responsive / unscorable.

5 / 5

Appendix F: HT RDF A-G Rubric and Instructions

HT RDF A-G Rubric and Scoring Instructions

Part 1 of Rubric: Identify expected elements of argumentation

Mark areas of text in the colors / letters noted below. Note some text will get multiple marks; and for example, often D or F examples will also be B examples; C and E are sometimes the same, and often a single sentence can do both A and B , or C and D, etc. The tag for a sentence can be a single character, or a string: A, or G, or "BD" or "CD" even "BDF".

A - Discuss how Harriet's key quality of leadership helped her overcome several obstacles and…

Tag a clearly stated "key quality" from the White paper (one of the seven). Picking one of the seven and identifying it clearly is good, adding explanations or other clarifying material could result in two "A" tags, but no need to tag repeated ways of making the same proposition.

B - (discuss) why [the key quality] was important to the her and the other slaves' survival. This is largely the same as citing evidence for the choice in A.

just tag a sentence if it identifies specific examples of, or evidence to reinforce the importance, of the chosen quality. Try to count each example or bit of evidence only once.

C - How is Harriet's response to life-threatening situation similar to how her followers react…

Tag sentences that show H and followers common responses. Some might also be things they have in common (E).

D - and how is Harriet's response to life threatening situations different from her followers' reactions?

Describing contrast between H vs. followers, or differential responses from H, to sever peril should be tagged with D. these might also explicitly be B or F items, depending on wording.

E - What does Harriet have in common with her followers; how is she similar / what she shares with them? Tag relevant sentences.
(e.g. they are all fearful for their lives; they are all discouraged by hunger / cold / privation; etc) Tag observations of commonality with E.

F - what differences allow Harriet to emerge as a leader? Tag each sentence.

(e.g. Harriet's inner motivation / passion / dedication; here encouragement + inspiration to others. Her sense of duty;

G - Conclusion: does it describe a lesson from Harriet's story and her acts of courage? A generalized conclusion or explicit lesson learned should be tagged with G.

1 / 4

For part one of this rubric, each portion of a response (sentence or part of a sentence) that satisfies some element of the above rubric shall receive a score of either 1 or 2 points. If portions of text span more than one sentence represent a desired response element, the point can be assigned to any sentence that is part of the relevant text.

[Note: As a simplifying assumption for our proof-of-concept implementation, feedback and scoring support will refer to an underlying sentence, and not a specific set of words.]

Part 2 of Rubric: Sum scoring micro-decisions to produce raw score total

Sum points assigned for all instances of each category identified in each response, limiting the category totals to a maximum value per category as specified below:

Caption: Maximum points per response element category

Category	A	B	C	D	E	F	G
Max Points	2	2	2	2	1	2	1

Individual response elements that satisfy rubric elements will therefor receive 1 or 2 points, and for each category, no more than 2 points (and in two cases only 1 point) will be included in the overall raw score.

Note that in addition to a maximum points-per-category, duplicate citations of the same evidence, or support for a given label (e.g. three statements elaborating on the "claim" or category 1 that are redundant), should not be counted. That is, each point awarded should be for a distinct contribution.

To complete part 2, sum the capped per-category total values to produce a total raw score, which will range between zero and a maximum of 12.

The figure below shows the scoring for a single response, and that an abundance of evidence results in a "capped" overall raw score total.

Figure 1 - Example of Capped Raw Score for Harriet Item using the A-G Argumentation Rubric

					max two	max two	max two	max two	max one	max two	max one		
seqn	irsen_id	itemresp_id	irsen _seqn	irpara _seqn	irs_a	irs_b	irs_c	irs_d	irs_e	irs_f	irs_g	(sum)	irsen_text
1	19365	3544	1	1	0	0	0	0	0	0	0		"Go on with us
2	19366	3544	2	1	0	0	0	0	0	0	0		Even after suc
3	19367	3544	3	1	0	0	0	0	0	0	0		Tubman's brav
4	19368	3544	4	1	2	0	0	0	0	0	0	2	Tubman's mos
5	19369	3544	1	2	0	0	0	0	0	0	0		Harriet Tubma
6	19370	3544	2	2	0	0	0	0	0	0	0		For instance w
7	19371	3544	3	2	0	1	0	0	0	0	0	1	It is better to b
8	19372	3544	4	2	0	0	0	1	0	0	0	1	She grimly saic
9	19373	3544	5	2	0	0	0	0	0	0	0		Tubman was fa
10	19374	3544	6	2	0	0	0	0	0	1	0	1	She didn't both
11	19375	3544	7	2	0	0	0	0	0	0	0		In the midst of
12	19376	3544	8	2	0	0	0	0	0	1	0	1	However, Tubı
13	19377	3544	9	2	0	1	1	0	1	0	0	3	Tubman, like a
14	19378	3544	1	3	0	0	0	0	0	0	0		Another exam
15	19379	3544	2	3	0	0	0	1	0	1	0	2	And freedom's
16	19380	3544	3	3	0	0	0	0	0	0	0		Those words a
17	19381	3544	4	3	0	0	0	0	0	0	0		As one of her f
18	19382	3544	5	3	0	0	0	0	0	0	0		"She tried to e
19	19383	3544	6	3	0	0	0	1	0	0	0	1	If a runaway re
20	19384	3544	7	3	0	0	0	0	0	1	0	1	These differen
21	19385	3544	1	4	0	0	0	0	0	0	0		"A leader with
22	19386	3544	2	4	0	1	0	0	0	0	0	1	People look to
23	19387	3544	3	4	0	0	0	0	0	0	0		Tubman's orga
24	19388	3544	4	4	0	0	0	0	0	0	1	1	A lesson that e
			raw score		2	3	1	3	1	4	1	15	
			final score		2	2	1	2	1	2	1	11	
					max two	max two	max two	max two	max one	max two	max one	max 12	
					a	b	c	d	e	f	g	tot	

Note that in the scoring figure above a total of 4 points were earned for citations of the "F" factor, or "what differences allow Harriet to emerge as a leader", but this raw score of 4 is reduced to 2 before the total raw score is calculated as 11 of 12. In this illustration, additional points for factors B and D were capped, lowering what might have been a raw score of 15 to a score of 11.

Part 3 of Rubric: Determine final score

The final score will be assigned to a value equal to one half the total raw score, rounding up as shown in the table below.

Raw Score Range	Final Scaled Score	Final Score Descriptor
8 - 12	3	Commendable / Exceptional Achievement. Response contains many or all of the suggested elements reflecting a good understanding of the question and the related texts.
5 - 7	2	Some / Adequate Achievement. Response contains an answer having some of the suggested elements, reflecting a basic understanding of the question and the related text.
1 - 4	1	Minimal / Little Evidence of Achievement. Response contains minimally relevant information reflecting at best a partial understanding of the question or of the texts.
0	0	No evidence of understanding the question or the related texts.

Appendix G: Rater Participant Information Form

Project Title: A Robust and Generalisable Rubric Design Framework for Critical Thinking Assessment

Invitation

You are being asked to take part in a research study on the development and validation of a rubric design framework for critical thinking assessment. The research is a PhD project conducted by Harry Layman under the supervision of Professor Glenn Fulcher at the University of Leicester.

What will happen

In this study, you will be given an item description, a rubric, and a set of responses. Some of the responses will be scored as examples. Unscored items are to be scored by you by applying the supplied rubric to the response provided. You will be provided with an information sheet detailing procedures, noting that your participation is voluntary, that you will be compensated for your time and at what rate, and explaining the means by which you may revoke your consent and discontinue your participation. Having agreed to participate in the study by signing the Consent Form below, you will be asked to rate or grade student responses to critical thinking challenges according to a rubric provided. You will also be asked to respond to a questionnaire following the rating sessions, answering questionnaire questions related to scoring process.

Time Commitment

Your time commitment will vary. You will be offered assignments with specific numbers of item responses to score and an expected (estimated) timeline for completion. Progress (scoring results so far) should be provided to the researcher after the first 20 items have been scored, and once approved, at least once every 40 hours of work, or 20 items scored, and compensation will be paid for each 40-hour interval or part thereof.

Participants' rights

You may decide to stop being a part of the research study at any time without explanation. You have the right to ask that any data you have supplied to that point be

withdrawn. You will be compensated for time spent for which scores have been provided. You have the right to omit or refuse to answer or respond to any question that is asked of you without any penalty.

You have the right to have your questions about the procedures answered (unless answering these questions would interfere with the study's outcome). If you have any questions as a result of reading this information sheet, you should ask the researcher before the study begins.

Benefits and risks

You will get some insights into approaches towards the evaluation of students' critical thinking responses. And you will also be invited to review the scoring results and the automated scoring platform in development. There are no known risks for you in this study.

Cost, reimbursement, and compensation

Your participation in this study is voluntary. You will receive compensation stated on your hourly rate as specified on the Rater Information Form for the hours of scoring work performed after each 40 hours worked, or when the work is completed, to compensate you for the time taken on your part.

Confidentiality / anonymity

The data to collect from you will include the results of your ratings, and your responses (if any) to questionnaires, name, email address, years of teaching or instructional experience, and years of assessment rating experience. Your name and all other information will be collected for identifying or indexing the rating results and questionnaire responses for the convenience of data analysis. Your email will be only used for the convenience of contact for the purposes of the research. In the presentation and publication of the research where your rating and questionnaire data are utilized, every precaution will be taken to protect your anonymity. This includes using pseudonyms; real names of individuals and universities will not be disclosed. All possible use of the data you provide will be only available to the researcher and the supervisory team under the above-mentioned conditions. Under all foreseeable

conditions, all use of the data will abide by the Data Protection Act 1998 and EU General Data Protection Regulation.[10]

For further information

Harry Layman / Professor Glenn Fulcher will be glad to answer your questions about this study and the final results of this study at any time. You may contact them at email: hal4@leicester.ac.uk /mobile: +1(949) 945-3373 and Dr. Fulcher email: gf39@le.ac.uk.

[10] See https://gdpr-info.eu

Appendix H: Rater Informed Consent Form

Project Title: A Robust and Generalisable Rubric Design Framework for Critical Thinking Assessment

Project summary

The project aims to develop and validate a rubric design framework to evaluating and improve rubrics for items designed to assess critical thinking skills with constructed response challenges or questions.

This project will have human scorers grade CT item responses with both generic, holistic rubrics and with item-specific, content-centric rubrics, to establish the relative IRR, utility and efficacy of each form of rubric in practice with the same assessment items and responses. Instructions, scoring materials and item responses and scoring forms will be provided.

By signing below, you are agreeing that: (1) you have read and understood the Rater Participant Information Form, (2) questions about your participation in this study have been answered satisfactorily, and (3) you are taking part in this research study voluntarily (without coercion)

Participant's Name (Printed)*

Participant's signature* Date

Name of person obtaining consent (Printed) Signature of person obtaining consent

For purposes of compensation: Compensate graduate research assistance for their time.

Hourly Rate or fixed milestone payments as specified on UpWork.com tasking. US and UK / EU graders will be compensated at a rate of no less than $15 / hour and €15, respectively.

Grader Name: _____

Grader Address: _____

Phone: Email:

_____ _____

Graders may have the option to register and use the UpWork platform for scoring work and payment.

College and University Education:

Institution	Years Attended	Major / Course of Study	Degree Awarded

Appendix I: Post Scoring Rater Survey

After scoring, raters received a questionnaire, one for each scenario they scored. The questions were focused on ease of use and the secondary research questions in this study regarding how the structured rubric impacted the scoring challenge from the perspective of the scorers. The three questions were applied to each scenario as illustrated by the survey instruments below.

Winter Hibiscus, C + Rubric Questions:

I am interested to in understanding how the rubrics may have made the scoring more or less efficient and more or less difficult.

1. The Winter Hibiscus with Claim plus Evidence rubric included specific guidance on partial scoring of the response claim. Was this:

Too specific: __ Too: broadly framed: __ About right: __

Comments:

2. The Winter Hibiscus with Claim plus Evidence rubric included specific guidance on scoring evidence to support the claim. Was this:

Too specific: __ Too: broadly framed: __ About right: __

Comments:

3. Did you find the rubric(s) easy or difficult to apply?

__ Easy __ Neutral __ Difficult

Comments:

4. Other Comments:

Harriet Tubman – C+E Rubric Questions:

1. The Harriet Tubman with Claim plus Evidence rubric included some guidance on partial scoring of the response claim. Was this:

I am interested to in understanding how the rubrics may have made the scoring more or less efficient and more or less difficult.

Too specific: __ Too: broadly framed: __ About right: __

Comments:

2. The Harriet Tubman with Claim plus Evidence rubric included guidance on scoring evidence to support the claim. Was this:

Too specific: __ Too: broadly framed: __ About right: __

Comments:

3. Did you find the rubric(s) easy or difficult to apply?

__ Easy __ Neutral __ Difficult

Comments:

4. Other Comments:

Harriet Tubman – A-G (Narrative Elements) Questions:

I am interested to in understanding how the rubrics may have made the scoring more or less efficient and more or less difficult.

1. The Harriet Tubman with A-G narrative elements rubric included guidance on partial scoring of the response claim and the reason for the claim (elements A and B). Was this:

Too specific: __ Too: broadly framed: __ About right: __

Comments:

2. The Harriet Tubman with Claim plus Evidence rubric included categorical guidance on scoring narrative elements C through G (e.g. how was HT's reaction to life threatening situations different from that of her followers; what lesson can we learn from HT's leadership, etc.) evidence to support the claim. Was this:

Too specific: __ Too: broadly framed: __ About right: __

Comments:

3. Did you find the rubric(s) easy or difficult to apply?

__ Easy __ Neutral __ Difficult

Comments:

4. Other Comments:

Appendix J: An RDF for CT and AW Scoring

A Rubric Design Framework (RDF) for CT and AW Assessment Items	
El #	Element Name / Description
Foundational Elements	
1	High-Level Rubric Definition
2	High-Level Item(s) Definition
3	Scoring Criteria and Level Definitions
4	Subscale Score Calculation Formula
5	Final Raw Score Formula
6	Score Scaling Formula
Scoring Processes	
7	Scoring Processes Strategy and Design
8	Scoring Process Implementation
Supporting Elements for Production Use	
9	Format and Content of Score Reports
10	Exemplars (example scored responses)

1. High-Level Rubric Definition
a) Construct: skill, knowledge or capability measured; Sub-scores and relative weights
b) Audience
c) How assessed
d) How scored
e) Security / disclosure

2. Item Definition
a) Instructions
b) Prompt / Challenge
c) Passage 1
d) Other artefacts (chart, graph, table, passage, sound, movie, image, interactions)

One row per subscore

3a. Scoring Criteria
a) Evaluative Criteria Name / Descriptor / Subscore 1
b) Evaluative Criteria Name / Descriptor / Subscore 2
c) Evaluative Criteria Name / Descriptor / Subscore 3 and so on…

One table per subscore

3b. For Each Evaluative Criteria: Quality Level Descriptors and Definition		
Subscore 1: Criteria Name		
Quality Level Name	Points	Definition
Level Name 1		Criteria by which this level is assigned, can range from categorical description to item specific detail
Level Name 2		
Level Name 3		
Level Name 4 and so on…		

One formula per subscore

4. Subscale score calculation formula
This entry expresses the rule by which a score for a subcategory is determined when one or more quality level definitions are satisfied for a single subscore or evaluative criteria.. Common rules might include: a) Adding up the qualifying individual point values of all the quality level definitions that are met by the response b) Selecting the highest point value from one or more quality level definitions that are met by the response c) Sum the point values of all satisfied quality level definitions that are met by the response up to a subscore maximum d) An algorithm specific to the item and the evaluation criteria (e.g., as might be used in a diagnostic radiology assessment or architectural engineering challenge)

5. Final raw score formula
If a single overall score is to be provided, this rule specifies how the final score is determined from the individual subscores. Common rules might include: a) Adding up subscores b) Averaging the subscores c) Summing the subscores up to a maximum possible

6. Final scaling formula
This formula specifies how the final raw score might be transposed from the calculated value to a reported target scale for comparability (across time, settings, or modalities) or for historical or other purposes based on external factors. Examples include a) A simple linear transformation. For example, a raw score of 0 to 60 could be transformed to a number between 200 and 800 in unit-of-10 increments; or b) A simple table could bracket raw scores into groups or categories of skill levels based on external validation or other processes. For example:

Raw score range	Final score	Final score descriptor
13–22	3	Strong evidence of recognising and understanding the central underlying analogy of the text.
8–12	2	Some evidence of recognising and understanding the central underlying analogy of the text.
1–7	1	Minimal evidence of recognising or understanding the central underlying analogy of the text.
0	0	No evidence of recognition or understanding the central underlying analogy of the text.

7. Scoring Process Strategy and Design
a) how will scoring be done; how scoring decisions be made and recorded
b) how scoring data is used to produce a score report
c) how meaningful feedback is produced
d) how scoring consistency and quality will be maintained

8. Scoring Process Implementation
Beyond classroom-based and instructor led scoring, in the case of large-scale assessment, consequential or high-stakes testing, and assessments for purposes beyond the focus of individual progress and achievement, scoring process implementation may require additional documentation to insure quality, validity, and reliability. Validity for these purposes may require more extensive work to provide assurance that measurement and score representations adequately reflect knowledge, skills and capabilities appropriate for the intended use of the assessment for all sub-populations and cohorts. Equivalence across time, geography, and populations requires an additional level of validation and documentation that may not be necessary for other contexts (e.g,. formative assessment, in-classroom assessment, or self-study).

9. Format and content of score reports
Some assessments may require different forms of score reporting for different audiences. Depending on context and use case, the intended use of an assessment item may provide a quick snapshot or benchmark to compare with prior and future like assessments to track progress and identify situations for further review, or the focus may be on diagnostic output that provides detailed subscore and item response-specific context to elaborate on the correct and incorrect thought processes, reasoning, and analysis performed by the examinee. A short summary that sets expectations for the expected score report, or a sample with explanations of the sort of feedback and score scale to be used, also contributes to a fair and open assessment process.

10. Exemplars
A library of well-scored examples can be an efficient way to communicate the standards with which a variety of quality level definitions can be met by different responses to complex CT or AW constructed response item. Well scored examples can also provide guideposts and underpin narrative descriptions of what a comprehensive challenge to a critical thinking item should contain. They can also provide examples of contrasting approaches to the same problem and illustrate a range of rhetorical techniques that can prove useful in the CT and AW domains.

References

Adèr, H. J. (2008). Advising on research methods: A consultant's companion. Johannes van Kessel Publishing.

Adey, P. and Shayer, M. (1994) *Really raising standards: cognitive intervention and academic achievement.* London: Routledge.

Akmam, A. *et al*. (2018, April). 'Influence of learning strategy of cognitive conflict on student misconception in computational physics course', (from the 2nd International Conference on Mathematics, Science, Education and Technology, Padang, Indonesia, 5–7 October 2017), *IOP Conference Series: Materials Science and Engineering,* 335, pp. 1–7. doi: 10.1088/1757-899X/335/1/012074.

Alderson, C. (1991) 'Bands and scores', in Alderson, C., and North, B. (eds.) *Language testing in the 1990s: the communicative legacy.* London: Modern English Publications/British Council/Macmillan, pp. 71–86.

Association of American Colleges and Universities (2005). *Liberal education outcomes: a preliminary report on student achievement in college.* Washington, D.C.: Association of American Colleges and Universities.

Association of American Colleges and Universities (2011) *The LEAP vision for learning: outcomes, practices, impact, and employers' view.* Washington, D.C.: Association of American Colleges and Universities.

Bailin, S. (2002) 'Critical thinking and science education', *Science & Education,* 11, pp. 361–375. doi: 10.1023/A:1016042608621.

Baldwin, D., Fowles, M. and Livingston, S. (2005) *Guidelines for constructed-response and other performance assessments*. Princeton, N.J.: Educational Testing Service. Available at: https://www.ets.org/Media/About_ETS/pdf/8561_ConstructedResponse_guidelines.pdf (Accessed: September 19, 2018).

Barnet, S., Bedau, H.A. and O'Hara, J. (2008) *Critical thinking, reading, and writing: a brief guide to argument*. Boston, Mass.: Bedford/St. Martins.

Bean, J.C. (2011) *Engaging ideas: The professor's guide to integrating writing, critical thinking, and active learning in the classroom*. 2nd edn. San Francisco, Calif.: Jossey-Bass.

Bejar, I.I. (2017) 'A historical survey of research regarding constructed-response formats', in Bennett, R.E. and Von Davier, M. (eds.) *Advancing human assessment.* New York, N.Y.: Springer, pp. 565–633.

Bennett, R.E. (1991) *On the meanings of constructed response.* Princeton, N.J.: Educational Testing Service.

Birenbaum, M. (1996). 'Assessment 2000: Towards a pluralistic approach to assessment', in Birenbaum, M. and Dochy, F.J.R.C. (eds.) *Alternatives in assessment of achievements, learning processes and prior knowledge*. Dordrecht: Springer, pp. 3–29.

Brindley, G. (1991) 'Defining language ability: the criteria for criteria', in S. Anivan (ed.) *Current developments in language testing*. Singapore: Southeast Asian Ministers of Education Organization, pp. 139–154.

Butterworth, J. and Thwaites, G. (2013) *Thinking skills: critical thinking and problem solving*. Cambridge: Cambridge University Press.

Cambridge Assessment (no date) *Thinking Skills Assessment.* Available at: http://www.admissionstesting.org/for-test-takers/thinking-skills-assessment/ (Accessed: 1 May 2018).

Cambridge Assessment (2018) *BMAT Biomedical Admissions Test*. Available at: https://www.admissionstesting.org/Images/201905-bmat-guide-for-universities-and-policy-makers-.pdf (Accessed: 16 September 2018).

Chase, B.J. (2011) *An analysis of the argumentative writing skills of academically underprepared college students*. PhD thesis. Columbia University. Available at: http://academiccommons.columbia.edu/download/fedora_content/download/ac:131454/CONTENT/Chase_columbia_0054D_10083.pdf (Accessed: 30 April 2018).

Coirier, P., Andriessen, J. and Chanquoy, L. (1999) 'From planning to translating: the specificity of argumentative writing', in Coirier, P. and Andriesse, J. (eds.) *Foundations of argumentative text processing*. Amsterdam: Amsterdam University Press, pp. 1–28.

The College Board (no date) *Compare the SAT to the ACT.* Available at: https://collegereadiness.collegeboard.org/sat/inside-the-test/compare-new-sat-act (Accessed: 24 June 2020).

Council for Aid to Education (no date) *CLA+ for higher education.* Available at: http://cae.org/flagship-assessments-cla-cwra/cla/ (Accessed 1 May 2018).

Council for Aid to Education (2013) *CLA+ scoring rubric*. Available at: http://cae.org/images/uploads/pdf/CLA_Plus_Scoring_Rubric.pdf (Accessed 22 March 2016).

Cumming, A., Kantor, R. and Powers, D.E. (2002) 'Decision making while rating ESL/EFL writing tasks: a descriptive framework', *The Modern Language Journal,* 86(1), pp. 67–96. doi: 10.1111/1540-4781.00137.

Dawson, P. (2017) 'Assessment rubrics: towards clearer and more replicable design, research and practice', *Assessment & Evaluation in Higher Education,* 42(3), pp. 347–360. doi: 10.1080/02602938.2015.1111294.

Deane, P. *et al.* (2008) *Cognitive models of writing: writing proficiency as a complex integrated skill.* doi: 10.1002/j.2333-8504.2008.tb02141.x.

DeRemer, M.L. (1998) 'Writing assessment: raters' elaboration of the rating task', *Assessing Writing,* 5, pp. 7–29. doi: 10.1016/S1075-2935(99)80003-8.

Facione, P.A. (1990) *Critical thinking: a statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations.* Available at: https://eric.ed.gov/?id=ED315423 (Accessed: 3 March 2020).

Frey, B.B., Schmitt, V.L. and Allen, J.P. (2012) 'Defining authentic classroom assessment', *Practical Assessment, Research & Evaluation*, *17*(2). doi: 10.7275/sxbs-0829.

Graduate Management Admissions Council (2016) *Analysis of an argument* [task rubric]. Available at: http://www.gmac.com/~/media/Images/gmac/GMAT/archive/analysisofanargument.pdf (Accessed: 22 March 2016)

Gulikers, J. T., Bastiaens, T. J. and Kirschner, P. A. (2004) 'A five-dimensional framework for authentic assessment', *Educational Technology Research and Development,* 52(3), pp. 67–86. doi: 10.1007/BF02504676.

Halpern, D.F. (2010) *The Halpern critical thinking assessment: manual.* Modling, Austria: Schuhfried GmbH.

Halpern, D.F. (2014) *Thought and knowledge: an introduction to critical thinking* (5th edn.). New York, N.Y.: Psychology Press.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. Assessment in Education: Principles, Policy & Practice, 20(3), 281-307.

Jackson, T. R., Draugalis, J. R., Slack, M. K., & Zachry, W. M. (2002). Validation of authentic performance assessment: A process suited for Rasch Modeling. American Journal of Pharmaceutical Education, 66, 233–242.

Johnson, R.L., Penny, J. and Gordon, B. (2001) 'Score resolution and the interrater reliability of holistic scores in rating essays', *Written Communication,* 18(2), pp. 229–249. doi: 10.1177/0741088301018002003.

Jonsson, A. and Svingby, G. (2007) 'The use of scoring rubrics: reliability, validity and educational consequences', *Educational Research Review,* 2(2), pp. 130–144. doi: 10.1016/ j.edurev.2007.05.002.

Kane, M. (2010) 'Validity and fairness,' *Language Testing,* 27(2), pp. 177–182. doi: 10.1177/0265532209349467.

Kaulfers, W.V. (1944) 'Wartime development in modern-language achievement testing', *The Modern Language Journal,* 28(2), pp. 136–150. doi: 10.2307/317331.

Kelly, F.J. (1916) 'The Kansas Silent Reading Tests', *The Journal of Educational Psychology*, 7(2), pp. 62–80. doi: 10.1037/h0073542.

Kuhn, D. (1999) 'A developmental model of critical thinking', *Educational Researcher,* 28(2), pp. 16–46. doi: 10.2307/1177186.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. biometrics, 159-174.

Lazer, S. *et al.* (2010) *Thoughts on an assessment of common core standards.* Available at: https://www.ets.org/s/commonassessments/pdf/ThoughtsonAssessment.pdf (Accessed: 3 March 2020).

Lin, S.-S. (2014) 'Science and non-science undergraduate students' critical thinking and argumentation performance in reading a science news report', *International Journal of Science and Mathematics Education,* 12(5), pp. 1023–1046. doi: 10.1007/s10763-013-9451-7.

Liu, O.L., Frankel, L. and Roohr, K.C. (2014) *Assessing critical thinking in higher education: current state and directions for next-generation assessment* (ETS Research Report No. RR-14-10). Available at: https://www.ets.org/research/policy_research_reports/publications/report/2014/jsjg (Accessed 2 April 2018).

Lomask, M. S., & Baron, J. B. (2003). What can performance-based assessment tell us about students' reasoning? In: D. Fasko (Ed.), Critical thinking and reasoning current research, theory, and practice (pp. 331–354). Cresskill, NJ: Hampton Press.

Marzano, R.J. (2002) 'A comparison of selected methods of scoring classroom assessments', *Applied Measurement in Education,* 15, pp. 249–267. doi: 10.1207/S15324818AME1503_2.

McClellan, C.A. (2010) 'Constructed-response scoring—doing it right', *R&D Connections,* 13, pp. 1–7. Available at: https://origin-www.ets.org/Media/Research/pdf/RD_Connections13.pdf (Accessed: 3 March 2020).

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica, 22(3), 276-282.

Messick, S. (1980) 'Test validity and the ethics of assessment,' *American Psychologist,* 35(11), pp. 1012–1027. doi: 10.1037/0003-066X.35.11.1012.

Messick, S. (1989). 'Meaning and values in test validation: The science and ethics of assessment', *Educational Researcher*, *18*(2), pp. 5–11. doi: 10.3102%2F0013189X018002005.

Messick, S. (1994) 'The interplay of evidence and consequences in the validation of performance assessments', *Educational Researcher,* 23, pp. 13–23. doi: 10.3102/0013189X023002013.

Messick, S. (1995) 'Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning', *American Psychologist,* 50(9), pp. 741–749. doi: 10.1037/0003-066X.50.9.741.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. Applied Psychological Measurement, 24 (4), 367–378.

Myers, M. (1980) *A procedure for writing assessment and holistic scoring*. Urbana, Ill.: National Council of Teachers of English.

National Center for Education Statistics (2008) *NAEP technical documentation: scoring monitoring.* Available at: https://nces.ed.gov/nationsreportcard/tdw/scoring/scoring.asp (Accessed: 20 January 2020).

Nordrum, L., Evans, K. and Gustafsson, M. (2013). 'Comparing student learning experiences of in-text commentary and rubric-articulated feedback: strategies for formative assessment', *Assessment & Evaluation in Higher Education,* 38(8), pp. 919–940. doi: 10.1080/02602938.2012.758229.

OCR (2013) *Specification—AS/A level critical thinking*. Available at: https://ocr.org.uk/Images/73470-specification.pdf (Accessed: 14 September 2018).

Palm, T. (2008) 'Performance assessment and authentic assessment: A conceptual analysis of the literature,' *Practical Assessment, Research, and Evaluation,* 13(1). doi: 10.7275/0qpc-ws45.

Pascarella, E.T. and Terenzini, P.T. (2005) *How college affects students: a third decade of research* (Vol. 2). San Francisco, Calif.: Jossey-Bass.

Popham, W.J. (1997) 'What's wrong—and what's right—with rubrics', *Educational Leadership,* 55(2), pp. 72–75.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. The Journal of Technology, Learning and Assessment, 4(4).

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity or rubrics for assessment through writing. Assessing Writing, 15, 18–39.

Sadler, D.R. (2009) 'Transforming holistic assessment and grading into a vehicle for complex learning', in Joughin, G. (ed.) *Assessment, learning and judgement in higher education.* Dordrecht: Springer, pp. 45–64.

Saxton, E., Belanger, S., & Becker, W. (2012). The Critical Thinking Analytic Rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments. Assessing writing, 17(4), 251-270

Siegel, H. (1988) *Educating reason: rationality, critical thinking, and education.* New York, N.Y.: Routledge.

Smarter Balanced Assessment Consortium (2014). 4-point *argumentative performance task writing rubric (Grades 6-11)*. Available at: https://portal.smarterbalanced.org/library/en/performance-task-writing-rubric-argumentative.pdf (Accessed September 20, 2018).

Tierney, R. and Simon, M. (2004) 'What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels', *Practical Assessment, Research & Evaluation,* 9, Article 2. doi: 10.7275/jtvt-wg68.

Timmerman, B.E.C. *et al.* (2011) 'Development of a "universal" rubric for assessing undergraduates' scientific reasoning skills using scientific writing', *Assessment & Evaluation in Higher Education,* 36(5), pp. 509–547. doi: 10.1080/02602930903540991.

Toulmin, S.E. (2003) *The uses of argument*. Cambridge: Cambridge University Press.

Torrance, H. (2007) 'Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning', *Assessment in Education: Principles, Policy & Practice,* 14(3), pp. 281–294. doi: 10.1080/09695940701591867.

Wiggins, G. (1990) 'The case for authentic assessment', *Practical Assessment, Research and Evaluation,* 2(2). doi:10.7275/ffb1-mm19.

Williamson, D.M., Xi, X. and Breyer, F.J. (2012) 'A framework for evaluation and use of automated scoring', *Educational Measurement: Issues and Practice,* 31(1), pp. 2–13. doi: 10.1111/j.1745-3992.2011.00223.x.

Wolfe, E.W. (1997) *A study of expertise in essay scoring.* Unpublished PhD dissertation. University of California, Berkeley.

Wood, B. (1928). *New York experiments with new-type modern language tests*. New York, N.Y.: Macmillan.

Yeh, S. S. (2001). Tests Worth Teaching To: Constructing State-Mandated Tests That Emphasize Critical Thinking. Educational Researcher, 30(9), 12–17. https://doi.org/10.3102/0013189X030009012

Zahner, D. (2013) *Reliability and validity–CLA+.* New York, N.Y.: CAE.

Zhang, M. (2013) Contrasting automated and human scoring of essays. *R & D Connections*, *21*(2). Available at: https://www.ets.org/research/policy_research_reports/publications/periodical/2013/jpdd (Accessed: 3 March 2020).

Zinsser, W.K. (1988) *Writing to learn*. New York, N.Y.: HarperCollins.