Cavanaugh, Gordon (2016) Heuristic assessment : professional judgement and final report cards in Ontario secondary schools. PhD thesis, University of Nottingham.

Access from the University of Nottingham repository:
http://eprints.nottingham.ac.uk/31034/1/GWCFinalThesis.pdf

Copyright and reuse:

For more information, please contact eprints@nottingham.ac.uk

HEURISTIC ASSESSMENT:
PROFESSIONAL JUDGEMENT AND FINAL REPORT CARDS IN
ONTARIO SECONDARY SCHOOLS

GORDON WILLIAM CAVANAUGH

**Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy**

FEBRUARY 2016

ABSTRACT

This thesis explored teacher professional judgement as applied to the final report card process of Ontario Secondary School courses in Business, Humanities, and Social Science. A constructivist grounded theory approach was used. Twenty-four active teachers from various schools participated in semi-structured interviews and follow-up questions. How the respondents understood the use of professional judgement when determining percentage grades was analyzed. The study found that the participants personalized procedures, either independently or at the direction of the local administration, when interpreting policy into practice. These practices, although done with good intentions, were at odds with reliable and valid assessment. This phenomenon was termed Heuristic Assessment.

Ontario's revised assessment and evaluation policy *Growing Success* (Ontario, 2010a) placed emphasis on informed professional judgement. Although a definition was provided, how the concept works in practice was open to interpretation. Therefore, schools can apply professional judgement in numerous ways and still be in line with provincial policy if what is taught and evaluated correspond with curriculum documents. However, this study found that Ministry instructions are challenging to implement. There are tensions between how the local administration view policy, participant understanding of these guidelines, and the realities of the classroom. Furthermore, school culture consists of both shared, or public, and shadowed, or private practices. Shared and shadowed practices sometimes go with, and sometimes against, provincial policy. Consequently, participants engaged in Heuristic Assessment: they used their

professional judgement to adhere to local policy in appearance, while finding ways to evaluate final report cards on their own terms.

This study makes several contributions to the field of knowledge. First, we see the concept of professional judgement in Ontario evaluation practices not as an idealized definition but as teacher-created construct. Second, there was clear evidence that the province still has work to do in order to have better consistency in assessment of learning practices. Understanding gained by the research established proposals on how to further improve reporting of student learning in Ontario and other educational systems. For example, there are easier ways for teachers to explain the meaning of grades to students, parents, guardians, and other stakeholders. If professional judgement is vital to evaluation practices, then the concept should be reified to assist teachers with the assessment process. There is also a methodological contribution, as the study provided an example of how to blend the constructivist grounded theory of Kathy Charmaz with the situational analysis of Adele Clarke to educational evaluation research.

ACKNOWLEDGMENTS

First, I must thank the participants. Without their willingness to share their insights on how Ontario's assessment policy works in practice, the study would not have been possible. I sincerely hope the findings of this study, in some way, will improve current practices. Their contribution may help shape the direction we take from here.

Second, I thank my supervisors Dr. Ganakumaran Subramaniam and Dr. Wei Keong Too for their advice during the study. I appreciate their recognition of the challenges associated with working part-time on a thesis while living a distance from the campus while also working full time as a teacher.

Finally, I want to thank my friends for their encouragement over the years. My parents, Ann and Garry, also deserve a special thank you. I would need the space of another thesis to describe my gratitude for all your love and support.

TABLE OF CONTENTS

**CHAPTER TWO: METHODOLOGY AND METHODS –
CONSTRUCTIVIST GROUNDED THEORY…………………………..52**

LIST OF CHARTS

LIST OF TABLES

CHAPTER ONE: PROFESSIONAL JUDGEMENT IN ONTARIO'S SECONDARY SCHOOLS

1.1 Introduction

This chapter provides an overview of the study and the research questions. The reader is introduced to the aims and objectives of the study, how and why the research was conducted, and the participants involved. The organization of the project is explained. A detailed account of the province's assessment and evaluation policy, *Growing Success* (Ontario, 2010a), explicates the various aspects Secondary School teachers need to consider and disregard when determining a report card grade – which all revolve around the idea of professional judgement. A close reading of this information helps to illustrate what educators must interpret to implement assessment policy. This interpretation process established the foundation for the research design and specific research questions. Furthermore, some referencing to the report's finding are included. To summarize, this grounded theory study seeks to understand how the participants conceptualized professional judgement as part of their final report card procedures, and concluded that this discretion is personalized in what was termed Heuristic Assessment in order to make sense of a complex assessment situation. Heuristic Assessment demonstrates the problematic nature of assessment practices in Ontario, as real world factors generate challenges the classroom teacher must solve as part of informed professional judgement while also attempting to perform valid and reliable evaluations.

1.2 Research Inquiry and Rationale

In Ontario's Secondary Schools, all course assessments and evaluations – including the final report card – are conducted by the classroom teacher. Teachers examine a student's achievement, compare it to the content and performance standards, and express it as an overall percentage grade. According to provincial policy, central to this process is the concept of informed professional judgement. In practice, to the educators who apply it, what is meant by professional judgement in the final evaluation process? Policy provides some description of this term, but its application is left to the interpretation of educators.

Following a constructivist grounded theory approach, this study aimed to co-construct an explanation of professional judgement. Professional judgement is a vast concept. Consequently, the research began not with specific research questions, but with a general inquiry statement (i.e., what is professional judgement?). To narrow the scope of the research participants were asked to define professional judgement, and to provide examples of how it is applied to final report cards, to focus on how it manifested in the evaluation process. Furthermore, how the participants created this interpretation revealed details of the social world of Ontario's schools, the arenas such as teacher-student interactions, and grading as a form of negotiation.

The objective of this study was to provide insight on a common term that is commonly misunderstood. In other words, professional judgement is a familiar term to Ontario teachers, but its functional definition is somewhat elusive.

Constructivist thought asserts that research findings are an interpretation of a social process or phenomenon, and should not be taken as an all-encompassing rule. However, a well-crafted interpretation, grounded in data, can advance knowledge (Charmaz, 2014). Since so much social meaning is generated by report card grades, and professional judgement is the key in determining these grades – this decision process should not be enigmatic. Therefore, this study established a goal of making a contribution to the field of knowledge by discussing the multifaceted role of professional judgement in Ontario's final report card evaluations in the hope of generating further discussion of how to improve grading practices.

This study found that the application of professional judgement in Ontario schools is in need of a transparent conversation. The social world of evaluation in the province is comprised of competing and contradicting conventions. Granted, social worlds are complex and confusing by nature (Clarke, 2005). However, a generation of assessment and evaluation reform in Ontario was meant to establish consistency in the form valid and reliable grades. The intention of informed professional judgement is to guide educator prudence to associate a student's consistent achievement with a score based on communal standards. Unfortunately, educators still face a complicated system of shared and shadowed practices within individual schools. Consequently, how to best determine a student's grade is challenging on a number of levels. In order to simplify the process, professional judgement, as applied to the final report card, becomes a heuristic device on which to base an evaluation not only of student achievement, but takes local demands and

perceptions into account. Furthermore, the participants tended to err of the side of caution and slightly increased percentage grades mostly based on non-achievement factors. In short, the practice of Heuristic Assessment questions the notion of true assessment consistency in the province, not to mention issues associated with validity and reliability. Better understanding why the participants applied Heuristic Assessment could lead to greatly improved evaluation practices by explaining to educators why this approach to grading is inappropriate.

1.3 Statement of Personal Interest and Inspiration for the Study

In a constructivist study, the voices of the participants should be dominant. At the same time, as a co-constructor of knowledge, the researcher should not hide behind data. Disclosure regarding personal viewpoints and prior assumptions is important information for the reader to help determine the trustworthiness of the findings. Furthermore, these details need to be confronted to help the researcher avoid forcing meaning. The grounded theorist must balance presuppositions and data (Charmaz and Mitchell, 1996; Rennie, 2000; Caelli, Ray and Mill, 2003; Mills, Bonner and Francis, 2006a; Hoare, Mills, and Francis, 2012). The researcher can also reflect on how the data changes one's point of view (Delamont and Atkinson, 2010). Throughout this study (for example, see Sections 2.3.7.1, 2.3.7.5, 2.9.7.1.2, and 2.10.5.2), I have reflected on my role as a researcher and my personal connection to the study. A brief explanation of why I pursued this report should provide disclosure and help alleviate any concerns about being too close to the topic under investigation.

To tell the truth, I began my teaching career with no previous preparation on how to determine a report card grade; my training program focused on formative assessment. The only advice I received was to use my professional judgement when evaluating. However, the Ontario Ministry of Education (OME) had no definition for professional judgement, and administrators and colleagues offered various interpretations. An analytical journey to explore the connotative and denotative of professional judgement resulted in pursuing a PhD on the topic.

By the time the study began I had been through the final report card process several times, but professional judgement was still mysterious. I was often unsure how to determine a grade, and my colleagues had difficulty clearly explaining their own processes. There was relatively little in the literature on professional judgement, especially in regards to the situation in Ontario. Therefore, I proceeded to work on a research design that would allow me to look into how educators at other schools viewed professional judgement. This approach led me to the potential use of constructivist grounded theory. Its epistemology corresponds with my own view that knowledge is interactive and socially created. Its methodology was congruent with the research situation: a researcher with some prior knowledge wanted to know more about an area where limited previous work had been done, with the opportunity to co-create understanding directly with participants.

The June 2010 release of *Growing Success*, a short time into the study, emphasized the timeliness of the research. In this revised policy statement, not only was professional judgment given a prominent place but the OME provided a

definition. However, the application of professional judgement was still left to personal interpretation. This encouraged me to continue pursuing the investigation. *Growing Success* assisted the study by refining research and interview questions while helping me identify significant differences between policy and practice.

Throughout the research, I was legitimately surprised at the inconsistency of assessment and evaluation practices between schools, often linked to the misunderstanding or misapplication of professional judgement and its connection to evaluation policy. The research inspired me to share these findings within a conceptual framework compared to the extent literature. I hope my recommendations for changes to assessment practices in Ontario will bring positive change to educators and students alike.

1.4 The Participants

Due to the complexity of the concept of professional judgement, it was necessary to establish parameters regarding what aspects would be analyzed and discussed. Twenty-four active teachers of Business, the Humanities, and Social Science courses in Ontario's Secondary Schools agreed to discuss how they determine final report card grades. Because professional judgement emphasizes the subjective side of assessment and evaluation, it made sense to look at subjects where the curriculum itself is subjective. Furthermore, I am more familiar, as an Ontario educator, with these courses. It should also be noted that the respondents reflected on the evaluation of non-exceptional students without any modifications

to the curriculum expectations as per Ontario's procedures on Special Education. Section 2.10.1.1 explains the purposeful sampling method used to target respondents who would be knowledgeable of the topic area to help co-construct a working understanding of the process of professional judgement (Creswell, 2008). Section 2.10.1.3 also provides a profile of these respondents.

1.5 Organization of Study

As previously stated, this chapter provides an overview of the research. The next chapter will demonstrate how constructivist grounded theory, inspired by Kathy Charmaz (2006; 2014) guided the methodology and epistemological framework. To address certain postmodern concerns, situational analysis, as described by Adele Clarke (2005), was also used. A methodology informs the methods, such as how data is gathered and categorized to establish insight. The different steps of the research process are accounted for and explained to demonstrate how the qualitative findings were validated.

In the third chapter, we move on to the analysis. It is shown how the voices of the participants answered the various questions raised in this study's Introduction. Categories cover the what, how, and why aspects of the research questions (Charmaz, 2008b). These categories contributed towards the core category of Heuristic Assessment that unveiled that, in practice, professional judgement guided the participants to individually adapt to their school environment, and balance the demands of provincial and local policies, in regards to report card evaluations. The respondents found ways to simplify the evaluation

process in a way that made sense to them, which involved using shared and shadowed practices. Professional judgement is not something that can be explained neatly and objectively, given the subjective nature of assessment itself – especially in a place as diverse as Ontario. Nevertheless, it is amazing how relatively straightforward heuristic approaches can be noted across the participants with various levels of experience in different schools, such as finding ways to justify higher grades.

The fourth chapter provides a literature review and discussion. Although the analysis helps to answer the research questions, bringing in other points of view supplements the findings. The present chapter includes some sources below for the purposes of policy discussion. The actual literature review will revisit these sources for additional clarity and connections with the analysis. Furthermore, examining the work of others assisted with making suggestions on how to improve assessment practices.

The final chapter proposes amendments to assessment policy and practices in Ontario. These suggestions, inline with the current generation of assessment reform, can make assessment and evaluation easier for educators to apply, and for students, parents/guardians to understand. Therefore, there would less likely be a discrepancy between shared and shadowed practices, thus vastly improving transparency. The problematic situation of Heuristic Assessment could be replaced with improved informed professional judgement. The proposals highlight the contributions this study has made to the field of knowledge by encouraging

Ontario educators to view their assessment procedures in a new light and promoting a revived way to view assessment and evaluation.

1.6 Forming the Research Questions: A Close Reading of *Growing Success*

The current standards-based educational reform in Ontario, also referred to as the New Curriculum, began in the mid-1990s (Anderson and Jaffa, 2003; Zegarac and Franz, 2007). Part of this reform was revamping the assessment and evaluation model. Several policy statements were released, most notably Ontario (1999a), (1999b), and (2000), but there were understandable obstacles with enacting significant changes in a multicultural province twice the size of France. For our purposes, it is not necessary to narrate a history of the Ontario's educational reform movement or the province's human geography. An excellent account of such details can be found in Gidney (1999). Those unaware of Ontario's educational makeup should note there are four major school boards (English Catholic, English Public, French Catholic, and French Public), comprising 72 individual boards. All schools, with tens of thousands of educators and over two million French and English students, including a significant indigenous and immigrant population, not to mention a strong heritage of school independence, are governed by the OME. Private Secondary Schools must also conform to Ministry policy. It should not surprise the reader that implementing centralized policies is a massive undertaking.

The current provincial government credited assessment reform as one of the reasons Ontario was recognized as one of the most improved school systems in

the world (Mourshed, Chijioke and Barber, 2010; Fullan, 2012; 2013). Still, it was

acknowledged that an improved and unified assessment policy was required to

further improve the student learning experience in Ontario. The result was the

release of *Growing Success: Assessment, Evaluation, and Reporting in Ontario's

Schools: Grades 1 to 12* (Ontario, 2010a). The document consolidated previous

assessment documentation in the hope of clarifying proper procedures for better

province-wide consistency. However, *Growing Success* did not establish any step-

by-step instructions, as the policy is meant to be flexible to suit the needs of

different boards. The concept of professional judgement, a term used in the past

but its actual role indeterminate, was upheld as the driving force behind best

practices. Therefore, to understand professional judgement, and how it is used as

part of the final report card evaluation process, one needs to investigate how the

concept is explained, understood, and applied.

The following subheadings are named after the relevant sections in

Growing Success. Each chapter of *Growing Success* is divided into two parts:

Policy and Context. Although the first part concerns the thinking behind the

policy, and the second elaborates on the application of ideas, the material in the

two sections tends to be repetitious. Therefore, this discussion of *Growing

Success* will not differentiate between the Policy and Context sections. Policy and

context, interpretation and action, are arenas within the situation of Ontario that

need to be explored in order to gain insight into the negotiation of professional

judgement. Noting the various aspects connected to professional judgement

helped to refine the overall investigation, and provides the reader with key aspects of the policy.

1.6.1 The Fundamental Principles

Growing Success stated Ontario policy is based on seven fundamental principles. Most importantly, assessment must be fair, transparent, and equitable. Policy also emphasized the importance of being consistent in assessment practices as "students and parents need to know that evaluations are based on evidence of student learning and that there is consistency in the way grades are assigned across schools and boards throughout the province" (p.2). Therefore, "students can have confidence in the information they use to make decisions about secondary pathways and postsecondary opportunities." Furthermore, colleges and universities, as well as employers, must know that final grades are based on common standards.

However, there are inevitable problems in implementing a unified assessment policy. *Growing Success* admitted:

> "Recognizing that the needs and circumstances of individual boards vary widely, the policy outlined in this document provides flexibility for boards to develop some locally focused guidelines and implementation strategies within the parameters for consistency set by the ministry. Education stakeholders throughout the province have voiced the need for greater consistency in assessment, evaluation, and reporting practices among the schools within a board, and initiatives to achieve improvement in that regard are strongly encouraged. Board guidelines should always be

developed in collaboration with all the schools in the board, and in

consultation with the school community." (p.2)

In order to accomplish the stated goals, the document explained, "Successful

implementation of policy depends on the professional judgement of educators at

all levels, as well as on educators' ability to work together and to build trust and

confidence among parents and students" (p.2). This is the first of the fourteen

times professional judgement is mentioned in the document. However, what is

professional judgement? According to *Growing Success*, professional judgement

is:

> "Judgement that is informed by professional knowledge of
> curriculum expectations, context, evidence of learning, methods of
> instruction and assessment, and the criteria and standards that
> indicate success in student learning. In professional practice,
> judgement involves a purposeful and systematic thinking process
> that evolves in terms of accuracy and insight with ongoing
> reflection and self-correction." (p.152)

In other words, it is a concept that involves taking knowledge of the Ontario

system and applying it in a methodical process and should become more refined

over time. The ramification is that much depends on the everyday educator, and

his or her decisions, to make this policy work. This definition will be explored to

better comprehend its connotation and denotation in practice throughout the study.

With this understanding, we will be able to better see how it actively blends the

different aspects of policy and practice.

Growing Success went on to claim that the Fundamental Principles help to "ensure that assessment, evaluation, and reporting are valid and reliable, and that they lead to the improvement of learning for all students" (p.6). In order for assessment to be valid, reliable, and assist learning, the importance of assessment to be transparent, equitable, and fair, was again emphasized. Since the document echoed these words, particularly in their connection to ensuring reliability, validity, and improving student learning, it would be worthwhile to examine the definition of each of these three terms.

In regards to transparency, *Growing Success* stated: "*transparency* is achieved when student learning is assessed and evaluated according to the clear standards outlined in the curriculum expectations (the content standards) provided in all curriculum documents… [and] outlined in the achievement chart that appears in every curriculum document" (p.7, emphasis in original). It should be clear that grades are objectively linked back to shared practices. Since both professional judgement and transparency have the common goal of upholding standards, their relationship with one another will be explored with participants. For instance, how do the participants compare achievement to both the curriculum expectations and the achievement chart transparently? Furthermore, how is this transparency communicated to students and other stakeholders? Answering these questions creates a fascinating discussion in Chapter Three.

Policy abruptly transitioned from transparency to a short discussion of equity, including details on the topic provided by the Organization for Economic Co-operation and Development (OECD, 2003). It simply stated that not all

students are to be treated the same. Students who require accommodations or modifications to the curriculum expectations must have their needs met, and assessed based on these provisions. The Glossary elaborated that equity was "A condition or state of fair, inclusive, and respectful treatment of all people. Equity does not mean that people are treated the same without regard for individual differences" (p.147). However, without expanding on equity and its important connection to assessment, a definition on fairness by Volante (2006, p.34) was inserted:

> "Fairness in assessment and evaluation is grounded in the belief that all students should be able to demonstrate their learning regardless of their socio-economic status, ethnicity, gender, geographic location, learning style, and/or need for special services." (cited in Ontario, 2010a, p.8)

In other words, students should have the same opportunity to be assessed, but are not necessarily assessed in the same way. Professional judgement is required to note this difference, but it is left to the reader to surmise the practical relationship between assessment, equity, fairness, and professional judgement in the Ontario situation. Furthermore, how these factors relate in a way that are consistent, valid, and reliable also not explained, thus starting a general theme of individual interpretation of the policy.

Growing Success made five references to validity in regards to classroom assessment and four to reliability. One needs to consult the Glossary to obtain definitions for these two key terms. First, reliability is defined as:

14

"The degree to which as assessment or evaluation is consistent and stable in measuring what it is intended to measure. An assessment or evaluation is considered reliable when the same results occur regardless of when and where the assessment or evaluation occurs or who does the scoring." (p.153)

On the other hand, validity is: "The degree to which an assessment or evaluation actually measures what it claims to measure and the extent to which inferences, conclusions, and decisions made on the basis of the results are appropriate and meaningful" (p.156). Again, policy mentioned these terms in passing, but it is not emphasized on how they are related to matters such as The Fundamental Principles or professional judgement. The concepts of reliability and validity will never be far from our discussion. Understanding the multiplicity of meaning of these terms, and the lack of conversation of their role in evaluation in Ontario, is essential to appreciate the phenomenon of Heuristic Assessment.

Growing Success went on to declare:

"Teachers have a leading role to play in the implementation of the seven fundamental principles. On a daily and hourly basis, teachers make professional judgements that ensure effective implementation of these principles, making decisions with respect to individual students and groups of students that have profound implications for them... In their important professional role, teachers show students that they care about them, and model a love of learning that can deeply influence their lives. Teachers' professional judgements are at the heart of effective assessment, evaluation, and reporting of student achievement." (p.8)

In this passage, we see how central the concept of professional judgement is to Ontario's assessment policy. It is the figurative heart of the system; active at all times. Whether or not a student feels engaged with the learning process, and perceives whether or not the educator "cares" about his or her progress, is essential to student success. How do the respondents react to this responsibility, and how is it expressed as professional judgement in regards to reporting student achievement? The first major parameter that needs to be examined is how the concepts of fairness, transparency, and equity influence professional judgement during the final report card process. Again, all discussion must also have a connection to assessment reliability and validity as we build a framework of Heuristic Assessment.

1.6.2 Learning Skills and Work Habits

The next section addressed Learning Skills and Work Habits Grades 1 to 12. Policy instructed educators not to include factors such as classroom behaviour (e.g., participation) and homework as part of the student's report card grade. Grades should be based on what the student has accomplished, not on the learning process itself. Skills are evaluated separately in the learning skills section of the report card, using the scale of Excellent, Good, Satisfactory, and Needs Improvement. Although this study does not address how the participants evaluated the learning skills, how the respondents perceived such skills is essential to our discussion. Since informed professional judgement guides the proper use of assessment, we can investigate the role of non-achievement issues when determining report card grades.

Growing Success listed the five learning skills which predate the revised document (which may have different labels depending on individual boards): Responsibility, Organization, Independent Work, Collaboration, Initiative, and added a sixth, Self-Regulation. Each contained bullet point descriptors. To condense the key aspects of each skill, in order: meeting deadlines; time management; completing tasks without constant teacher supervision; works well with classmates; taking the time to ask questions; and the qualities associated with assessment as learning (i.e., student metacognition). In other words, these are all positive behaviors that can contribute to student success. However, these behaviours are not to be confused with actual achievement.

The document explained, "the development of learning skills and work habits is an integral part of a student's learning," but *Growing Success* urged:

> "To the extent possible, however, the evaluation of learning skills and work habits, apart from any that may be included as part of a curriculum expectation in a subject or course, should not be considered in the determination of a student's grades. Assessing, evaluating, and reporting on the achievement of curriculum expectations and on the demonstration of learning skills and work habits *separately* allows teachers to provide information to the parents and student that is specific to each of the two areas of achievement." (p.10, emphasis in original)

The key phrase is "to the extent possible," suggesting that separating the evaluation of skills and grades can be challenging. The document went on to state:

> "In fact, achievement of the curriculum expectations in many curriculum areas is closely tied to learning skills and work habits.

Clearly identifying the focus of such curriculum expectations and the evidence that will be collected to assess and evaluate their achievement will assist teachers in making decisions about whether the demonstration of a learning skill or work habit should be part of the evaluation of a curriculum expectation." (p.10)

The document elaborated: "the development of the learning skills and work habits is further strengthened through the achievement of the curriculum expectations... These skills clearly overlap with and reinforce the learning skills and work habits… and will help students succeed in school and throughout their lives" (p.12). To clarify, in order to demonstrate achievement of the curriculum expectations, students utilize learning skills. An observant educator, applying professional judgement, should know where one stops and the other begins in regards to grading. At the same time, further developing these skills should help improve true achievement, which will further enhance student success as well as produce evidence of learning. When a participant reviews skills, how does this evidence enter into the thinking process that determines the overall grade?

The next major parameter we will look at is the role of student behaviour (i.e., non-achievement factors) in the grading process (see Sections 3.4 and 3.5). When behaviour and curriculum expectations blend so closely together, how do participants differentiate the two in a consistent manner, while also balancing the fundamental principles and curriculum expectations? Policy suggested that professional judgement makes it all possible, but what does it look like in practice?

1.6.3 Performance Standards – The Achievement Chart

Growing Success moved on to discuss the achievement chart and how it is connected to the provincial performance standards. The achievement chart is a means of providing transparency to students and parents/guardians regarding standards. It is the same template that has been in use since 1999 as part of the New Curriculum. Still, there are some issues to explore, as it is the professional judgement of the educator that navigates the chart, and connects it to standards, to justify grades.

Ontario teachers are to use both content and performance standards as part of their assessment practices. Content standards (also known as the curriculum expectations) direct what needs to be taught to students, and are broken down in to overall expectations and specific expectations. More precisely, students are evaluated on the overall expectations: a list of bullet points, separated by different themes, which are the governing topics, concepts, etc., of a course. The specific expectations are a detailed breakdown of the different aspects of the overall expectations, and includes examples of subject matter that could be taught. All specific expectations are to be covered as part of a course in order to prepare students for the evaluation of the overall expectations (Ontario, 2010a, p.28 and p.38). Report card grades are a statement of a student's consistent achievement of the overall expectations as directed by a teacher's professional judgement in relation to the performance standards.

According to *Growing Success*, the performance standards outlined on the achievement chart "enables teachers to make consistent judgements about the quality of student learning based on clear performance standards and on a body of evidence collected over time" (p.16). The performance standards establishes a template for the curriculum expectations, assists with the formation of rubrics, aids instructional planning, provides guidelines for comments on student achievement, and organizes the different learning categories by levels of achievement. This study is not about how the participants evaluate individual assignments; it is about how all these assignments come together to form the report card grade. Nevertheless, a discussion of the achievement chart helps to explain the formation of an overall grade.

The achievement chart is divided into learning categories. These categories (Knowledge & Understanding, Thinking & Inquiry, Communication, and Application) help stratify the curriculum into various streams to develop well-rounded students who not only have a comprehension of the course content, but are also able to unite the information, as well as communicate and apply it effectively. Just as all aspects of this policy were meant to work together, these "four categories should be considered as interrelated, reflecting the wholeness and interconnectedness of learning" (p.17). Moreover, educators should take a "balanced" approach to the learning categories, ensuring students have "numerous and varied opportunities to demonstrate the full extent of their achievement" (p.17). Educators are not required to teach the curriculum equally by learning category, but according to "*relative* importance" (p.17, emphasis in original).

Various assignments should be designed to overlap overall expectations in

numerous ways in order to strengthen and reinforce reliability and validity of

assessment, as well as to provide multiple chances to demonstrate attainment of

the expectations.

To decide on a report card grade, teachers need to reflect on the

combination of a student's curriculum accomplishments not only via the four

learning categories, but based on four achievement levels as well. Knowledge &

Understanding asks that the teacher rate the student's awareness of facts. The

other categories instruct the educator to reflect on the "*effectiveness*" of the

student's achievement, in regards to "clarity, accuracy, precision, logic, relevance,

significance, fluency, flexibility, depth, or breadth" (p.18, emphasis in original).

The four learning categories use similar key words and each level has a

corresponding percentage range: Level 1 (50%-59%) reflects "limited"

effectiveness; Level 2 (60%-69%) "some" effectiveness; Level 3 (70%-79%)

"considerable" effectiveness; and Level 4 (80%-100%) reflects a "high degree of"

or "thorough" effectiveness. All four categories and levels are relative to grade

level and stream (i.e., Academic, Applied, and Locally Developed for Grades 9

and 10; University, University/College, College, and Workplace for Grades 11 and

12; all grades have Open courses that are available to all students) of the

curriculum expectations. However, the policy emphasized, a Level 4 "*does not*

mean that the student has achieved expectations beyond those specified for the

grade/course" (p.18, emphasis in original).

The above raised a plethora of questions. For instance, what if a student's work actually does exceed the grade level? Is the distinction between key words such as "considerable" and "thorough" subjective in practice? Furthermore, what do the participants think of Level 4 being a 20% range, while the rest are 10%? How did respondents interpret the chart's descriptors, while also balancing the learning categories and the needs of individual learners as part of their professional judgement, to determine the report card grade? In addition, how are the levels on individual assignments combined into a percentage grade in a fair, transparent, and equitable manner? Such questions inform much of our discussion, particularly in Section 3.5.

Next, *Growing Success* addressed criterion-referenced assessment. Criterion-based referencing, which replaced normative-based referencing as part of the New Curriculum, allows teachers to judge students using the achievement chart. *Growing Success* explained:

> "In the past, assessment and evaluation performance standards varied from teacher to teacher and from school to school, and this led to results that were not always fair for all students. Criterion-referenced assessment and evaluation ensure that the assessment and evaluation of student learning in schools across the province are based on the application of the same set of well-defined performance standards. The goal of using a criterion-based approach is to make the assessment and evaluation of student achievement as fair, reliable, and transparent as possible." (p.19)

Certainly criterion-referencing on a common chart offers a better chance at upholding the fundamental principles across the province, but what does "as

possible" suggest? Also, the term reliable is again used without clarification.

Since it is individual educators internally deciding on assessments, what systems

are in place to "ensure" that there is consistency, while allowing for professional

judgement? Furthermore, how did participants, especially when new to the

profession, know what individual students are capable of without, at least to a

limited degree, considering how all students in the class performed? What sources

do new educators tend to draw upon to help them learn to assess? In Sections such

as 3.5.6, we will see there are many legacy issues still influencing grading in

Ontario.

To recap, Ontario educators base instruction on curriculum expectations.

All specific expectations are delivered through daily lessons in order to support the

overall expectations. Both sets of expectations are driven by a balanced delivery

of the learning categories. Teachers use the criterion-referenced achievement

chart and its levels to evaluate performance of the expectations. These levels are

then combined and converted into an overall percentage grade. Although the

learning skills aid the learning categories, thus intertwined into the content

standards, the teacher must be mindful to separate capabilities from achievement

when evaluating the overall expectations. This study intends to explain how all of

this results in the end product of a report card percentage grade. However, we

have yet to discuss the different forms of assessment and how they relate to

professional judgement.

1.6.4 Assessment *for* Learning and Assessment *as* Learning

The next section was dedicated to the two types of formative assessment: assessment "for" learning and assessment "as" learning. Although this study did not analyze formative assessment per se, it is part of our conversation. After all, proper professional judgement demands that educators understand the different forms of assessment to apply discretion appropriately. Furthermore, Ontario is one of the many educational systems that have placed much faith in formative assessment (Bennett, 2011). This section also repeated the message to eschew behaviours influencing grades. Instead, how a student performs in class should be used to help guide formative assessment. At the same time, just as learning skills blend in with, but are separate from, evaluation, the chapter attempted to explain how to braid different forms of assessment to enhance student learning and professional judgement. It also raised the issue of whether or not evidence of learning obtained under formative circumstances can be used to determine a report card grade.

To monitor student process, teachers were encouraged to "gain assessment information" on a daily basis to monitor student progress (p.28). The document advised to collect assessment in multiple ways, including:

> "formal and informal observations, discussions, learning conversations, questioning, conferences, homework, tasks done in groups, demonstrations, projects, portfolios, developmental continua, performances, peer and self-assessments, self-reflections, essays, and tests" (p.28).

It is also stated that this evidence should be "triangulated," but does not explain how (p.34). Furthermore, readers were reminded that formative assessment tasks should not only balance the learning categories, but also thoroughly cover the specific expectations.

By checking on the progress of students, educators can be more confident in the their capabilities. Advice is offered on how teachers can include assessment as and for learning throughout the instructional period. Teachers should always observe evidence of student learning to find ways to help them improve, ideally by providing opportune comments and descriptive feedback. Also, students should be encouraged to assess the work of classmates as well as their own to reinforce understanding of the curriculum. Evidence from these interactions can be used to evaluate the learning skills as outlined in a previous section (pp.28-29). Furthermore, ample opportunity to practice achievement of the curriculum expectations should be provided before a student's knowledge and skills are tested. School administrations were advised to watch over all assessment practices to ensure they are done correctly "by encouraging continuing professional development among staff and by fostering a school-wide collaborative learning culture based on the sharing of knowledge and on a sense of collective responsibility for outcomes" (p.29). The participants were asked how the local administration monitors their assessment practices and how this mentoring impacts professional judgement (see Section 3.2).

The section continued with an excerpt from Sutton (1991):

"It is worth noting, right from the start, that assessment is a human process, conducted by and with human beings, and subject inevitably to the frailties of human judgement. However crisp and objective we might try to make it, and however neatly quantifiable may be our "results", assessment is closer to an art than a science. It is, after all, an exercise in human communication." (p.2, cited p.29)

Similar to the earlier Volante quotation, there is no elaboration of what the excerpt means. Furthermore, this section of the policy is on formative assessment, and the quotation sounds more like a definition of summative assessment, so the placement seems odd. Either way, how do the participants explain to students and/or parents/guardians that marking is "closer to an art than a science?" In other words, in cases where they need to explain their professional judgement, what do the participants say (see Section 3.5.5)? It is believed that by analyzing such details and by seeing how participants unravel policy, we can get better insight into professional judgement and into the assessment culture of Ontario's secondary schools.

Growing Success took a moment to further define its assessment concepts and terms. For example, "*assessment* is used to mean a set of actions undertaken by the teacher and student to gather information about student learning" (p.30). The document elaborated that terms such as diagnostic, formative, and summative assessments, are now more commonly referred to as assessment as, for, and of learning, respectfully. Research from Harlen (2006, p.104) was included:

"Using the terms "formative assessment" and "summative assessment" can give the impression that these are different kinds of assessment or are

linked to different methods of gathering evidence. This is not the case; *what matters is how the information is used*. It is for this reason that the terms 'assessment for learning' and 'assessment of learning' are sometimes preferred. The essential distinction is that assessment for learning is used in making decisions that affect teaching and learning in the short term future, whereas assessment of learning is used to record and report what has been learned in the past." (cited in, and emphasis added, by *Growing Success*, p.30)

This information rephrased what has already been stated: educators must use their professional judgement to decide how to best address the needs of the individual learner to foster success, while logging true achievement to contribute towards an evaluation of the overall grade. However, how do the participants navigate this subtle but significance difference in assessment practices? Moreover, it was not clearly stated in the policy whether or not teachers may count evidence for summative purposes if it was gained under formative circumstances. We will see how the participants negotiate the overall process of both assessment as, for, and of learning as part of our core category in Section 3.6.

An Ontario educator needs to note the everyday achievement of a student and incorporate the knowledge as part of informed professional judgement. Furthermore, professional judgement recognizes the difference between evidence of student learning for formative or summative purposes. This section of *Growing Success* stressed the importance of assessment as and for learning, and that overall student behaviour and student products such as homework should be used to guide the learning process and not to determine grades. However, how classroom observations and conversations could serve a summative purpose is not entirely

clear. Furthermore, statements on the subjective nature of assessment in general did not advise how to explain this reality to students, parents/guardians, and other stakeholders.

1.6.5 Evaluation

We now move on to the section entitled Evaluation. Given what was said in the previous section, it is curious the authors did not choose "Assessment of Learning." Nevertheless, the chapter provided an overview of the process teachers should use to create the report card grade. Not only did it include details about how to use summative assessment to determine consistent achievement with reliability and validity, it also included a discussion on the use of punitive measures. Professional judgement was again referred to as being essential, but how it actually operates in the real world was left open to individual interpretation. Therefore, a careful review of this chapter was needed for the purposes of this study. Additional subheadings that do not appear in the policy are used to assist with the organization of the present chapter.

1.6.5.1 Key Terminology

The reader was again reminded that "the primary purpose of assessment is to improve student learning" (p.38). Immediately following, it was pointed out the previous section was on assessment as/for learning and needed to be distinguished from the present section. Evaluation, assessment of learning, involves teachers making a value judgement, as represented by a symbolic marker, of student achievement based on the aforementioned achievement chart. Since all

assessment is meant to improve student learning, then assessment of learning, including the report card grade, should also have a formative purpose. After all, the only difference between formative and summative assessment is how evidence of learning is used. However, does the different assessment terminology result in confusion for the participants and/or their administrations?

In regards to reports cards, teachers must provide a symbolic representation of how well students have achieved the curriculum expectations at pre-specific periods during a Secondary School course. We will focus our discussion on the final report card. As previously mentioned, all specific expectations need to be taught and assessed formatively. Although each specific expectation contributes towards a theme as expressed by an overall expectation, and all overall expectations need to be subjected to assessment of learning multiple times in the course, not every specific expectation itself needs to be evaluated. "Teachers will use their professional judgement," the document explained, "to determine which specific expectations should be used to evaluate achievement of the overall expectations, and which ones will be accounted for in instruction and assessment but not necessarily evaluated" (p.38). Having the choice of specific expectations to evaluate is another example of making *Growing Success* pliable.

1.6.5.3 Using Achievement Evidence With Reliability and Validity

Similar to its approach for assessment as/for learning, *Growing Success* explained that assessment of learning was accomplished by collecting "evidence of student achievement for evaluation… from three different sources – *observations*,

conversations, and *student products*" over time. It added that, "using multiple sources of evidence increases the reliability and validity of the evaluation of student learning" (p.39, emphasis in original). This is a similar to what was said about formative assessment. As previously stated, policy did not elaborate on whether or not an educator can use evidence gathered under formative circumstances for summative purposes. If the evidence would benefit a report on the student's learning, then it would seem fair and equitable. However, how the student is informed raises issues of transparency. Furthermore, what happens to reliability and validity when various examples of student learning are mixed? The participants will explain how they balance such matters in practice throughout Chapter Three, and this information will be compared to the literature (see Section 4.7).

Policy did not explain how multiple pieces of evidence gathered at different times make evaluation more reliable and valid. If an educator hopes that student achievement improves over time, what does this say about reliability? Of course individual assessments should be designed so that there is a reliable correlation between achievement and the provincial standards, but how is this done consistently between educators and different learners? At the same time, can assessments or evaluations be valid, even if reliability is questionable? An analysis of the participant responses can address the question and provide a deeper understanding of the function of informed professional judgement (see Section 3.5).

1.6.5.3 Organizing Summative Assessments

The section on Evaluation provided on overview of the types of assignments that should be used for assessment of learning:

> "'Student products' may be in the form of tests or exams and/or assignments for evaluation. Assignments for evaluation may include rich performance tasks, demonstrations, projects, and/or essays. To ensure equity for all students, assignments for evaluation and tests or exams are to be completed, whenever possible, under the supervision of a teacher." (p.39)

The last line is interesting. Although tests and exams are usually proctored, assignments such as rich assessment tasks are usually completed outside the classroom. At the same time, "assignments for evaluation must not include ongoing homework that students do in order to consolidate their knowledge and skills or to prepare for the next class" (p.39). Again, the fine line is, students must be notified of which tasks will be count towards the grade, and what is for formative purposes only. This is all done in the interest of transparency, but how does it work in practice? Do students work as diligently on homework as they do on evaluated tasks? When determining the final report card percentage grade, do participants tend to reflect on the regular completion of homework, or did they concentrate solely on summative tasks (see Section 3.4.3)?

1.6.5.4 Determining the Report Card Grade

The document moved into an important discussion about how to determine the report card grade:

"Teachers will take various considerations into account before making a decision about the grade to be entered on the report card. The teacher will consider all evidence collected through observations, conversations, and student products (tests/exams, assignments for evaluation). The teacher will consider the evidence for all the tests/exams and assignments for evaluation that the student has completed or submitted, the number of tests/exams or assignments for evaluation that were not completed or submitted, and the evidence of achievement that is available for each overall expectation for a subject in a particular grade or course. In addition, the teacher will consider that some evidence carries greater weight than other evidence; for example, some performance tasks are richer and reveal more about students' skills and knowledge than others. Teachers will weigh all evidence of student achievement in light of these considerations and will use their professional judgement to determine the student's report card grade. The report card grade represents a student's achievement of overall curriculum expectations, as demonstrated to that point in time. Determining a report card grade will involve teachers' professional judgement and interpretation of evidence and should reflect the student's most consistent level of achievement, with special consideration given to more recent evidence." (p.39)

This passage is similar to other parts of the policy we have read, but with additional information. It came somewhat closer to stating that evidence of learning gained under formative circumstances can be used for summative purposes (i.e., evidence of achievement that is available for each overall expectation), but again did not explain how doing so can be done with transparency. That said, an educator must also consider the weight of the evidence. For instance, a rich assessment task is more telling of an individual

student's achievement than a random classroom conversation or observation. The teacher, using informed professional judgement, must reflect on what the individual student accomplished consistently throughout the course.

Furthermore, it should be noted that the directive of most consistent and *more* recent achievement is a slight change from the previous policy of most consistent and *most* recent achievement. "Most recent" suggested that teachers had to place emphases on recent achievement. Now they can give "special consideration" to "more recent" evidence. However, this instruction could be interpreted in a different ways. First, how does one view more recent evidence differently than most consistent evidence of achievement? Also, how should an educator take improvement into account? More importantly, how do the participants balance of these directives? In Chapter Three, participants will take us through the report card process to see how they understand and summarize achievement. Again, noting the reliability and validity of such decisions is important in answering the research questions.

1.6.5.4 Determining versus Calculating Grades

Another important aspect of this passage is how percentage grades are determined as opposed to calculated. It is a teacher's professional judgement, along with mathematics, that determines the overall percentage grade (p.40). However, *Growing Success* did not advocate one particular type of calculation. In fact, no mathematical method was mentioned. In other words, in a policy document meant to guide educators on how to determine a report card percentage

grade, there were no instructions on how to accomplish this task. The document

did add: "Teachers will benefit from leadership by the principal to ensure that

there is a common understanding among all staff about the process for determining

the final grade. The principal will work with teachers to ensure common and

equitable grading practices that follow ministry policy and board guidelines"

(p.39). Generally speaking, how do local administrations interpret *Growing*

Success in regards to calculating and determining grades and advise the

participants (see Section 3.2.3)?

1.6.5.4.1 Course Work and the Final Evaluation

Further complicating the process is the fact the final grade is also broken

down into two major components. In the Ontario Secondary School system, the

final percentage grade is based upon 70% course work and a 30% final evaluation,

both assessed by the classroom teacher. This final evaluation often takes for the

form of an exam, but may also come in another form (e.g., a culminating project,

multiple tests and assignments, etc.), which "allows the student an opportunity to

demonstrate comprehensive achievement of the overall expectations of the course"

(p.41). However, the final evaluation cannot be considered as part of most

consistent and/or more recent achievement. Although it is often the last

assessment in a course, the 30% final evaluation is considered a separate entity

from the 70% course work, and informed professional judgement does not apply.

The final evaluation mark is assessed objectively, and added to the previously

determined course mark. Policy does not go into detail to explain the significance

of this point. For example, if a student received a course work grade of 50%, but

then managed to score 100% on the final evaluation, under policy, the final grade of that student can only be 65% (Gill, 2013a). However, what actually happens in practice? If a student's achievement on the final evaluation is out of sync with the most consistent achievement, is there any leeway? Clarification from the respondents was definitely needed, and is provided in Sections such as 3.2.4.

1.6.5.4.3 Converting Levels Into Percentage Grades

This section elaborated on how levels, including the addition of sublevels to official policy, are converted into report card grades (p.40). For Secondary School students, the levels assigned on the achievement chart have a corresponding percentage as seen on Table 1.1: Achievement Levels and Corresponding Percentage Grades. In other words, not only do educators need to reflect on the most consistent and more recent level of achievement, divided between four learning categories, which may not be equally weighted, the various evaluations to summarize are organized by sublevels – then converted to a 100-point scale. Furthermore, none of the sublevels are pegged to a specific percentage mark. For example, a Level 3 could be 73%, 74%, 75%, or 76%, based on a teacher's informed professional judgement. How does an educator decide, on a range of 73-76%, what percentage grade a student should get on a report card when their most consistent achievement is a Level 3? Can this process be explained in a way that only not reflects the curriculum expectations, but the fundamental principles as well? Also, if a Level 3 (70%-79%) reflects meeting the provincial standard, how do the sublevels work? Do they mean, more or less meeting, meeting, and almost exceeding? Furthermore, if Level 4 means

exceeding expectations, what is the difference between a student getting 80% and another getting 99%? Both are exceeding, but how can a student exceed significantly more than another? In regards to professional judgement, if an educator felt two students were a Level 2 overall, do the students get the same percentage grade, or can one get 64% and the other 66%? Since acceptance to post-secondary comes down to a student's Grade 12 percentage average, how do we know a student who is at 79% at one school actually achieved the curriculum expectations better than a student at another school who received 77%, even though they are both evaluated as 3+? All of these points have vital connections to reliability, validity, and professional judgement. Feedback from respondents is needed to provide elaboration in these matters and provides the narrative for Chapter Three.

Table 1.1: Achievement Levels and Corresponding Percentage Grades

Achievement Level	Percentage Mark Range	Achievement Level	Percentage Mark Range
4+	95-100	2+	67-69
4	87-94	2	63-66
4-	80-86	2-	60-62
3+	77-79	1+	57-59
3	73-76	1	53-56
3-	70-72	1-	50-52

1.6.5.4.4 Borderline Students

In regards to passing, if a student achieves at least 50% in a course, a credit is granted. This grade reflects a student who has demonstrated the bare minimum

of achieving the curriculum expectations. However, what is the demarcation between not demonstrating even limited achievement, and just falling short? *Growing Success* admitted that there is still work to be done regarding minimum marks. It stated: "Individual boards will work collaboratively with their school communities to determine the *lower limit* of the range of percentage marks below 50 per cent that teachers may record on the report cards of students in Grades 9 to 12. It is important that a consistent approach is adopted among all the schools of a board" (p.41, emphasis in original). What are the procedures when a participant has a student who is consistently scoring less than 50% on individual assignments? What percentage grade is recorded on the report card (see Section 3.2.5)?

It should also be noted that Grade 9 and 10 students who do not complete enough course work and/or the final evaluation, based upon a teacher's professional judgement, may have the code of "I" recorded on their transcript (p.42). Students with such "insufficient evidence" of learning may be able to take a Credit Recovery course, make up for the missing work, and obtain a credit without needing to repeat the course in its entirety. In the interest of full disclosure, Credit Recovery is not available for Grade 11 and 12 students and a percentage grade must be recorded. Since there is already so much material to address in regards to the application of professional judgement in determining report card grades, this study will not look at the code of "I" or Credit Recovery in depth. Nevertheless, it is important to point out Credit Recovery as part of the overall situation as it can impact a teacher's professional judgement and whether or not to assign a grade of 50% (see Section 3.2.5).

1.6.5.5 Applying Punitive Measures

The section on Evaluation moved on to a polarizing issue in Ontario: how to assess punitive measures on late, missing, and plagiarized assignments. From 1999-2010, Ontario teachers were instructed that such assignments were a behavioural issue, and mark deductions and the use of a mark of zero were forbidden. The revised position is, "it must be made clear to students that they are responsible for providing evidence of their learning within established timelines, and that there are consequences for cheating, plagiarizing, not completing work, and submitting work late" (p.42). Similar to instructions regarding report card grades of less than 50%, boards and schools were asked to work together to determine the best approach to encouraging academic integrity. The document advised boards to develop policies addressing the detection and prevention of cheating and plagiarism. Again, the document is trying to allow flexibility for individual boards, while providing guidance to help ensure consistency on a vital issue. *Growing Success* recommended: "Policies will reflect a continuum of behavioural and academic responses and consequences, based on at least the following four factors: (1) the grade level of the student, (2) the maturity of the student, (3) the number and frequency of incidents, and (4) the individual circumstances of the student" (p.43). In other words, in lieu of prohibiting the use of punitive measures, the document leaned towards leniency by encouraging boards, and individual teachers, to consider the whole picture on a case-by-case basis. There is also seventeen bullet points on how to address academic

dishonesty, with the last point being punitive measures, hinting they were to be used as a last resort.

Educators were advised to use their "professional judgement" for the best course of action in regards to late, missing, and plagiarized assignments (p.43). Most importantly, it must be ensured "that mark deduction will not result in a percentage mark that, in the professional judgement of the teacher, misrepresents the student's actual achievement." In addition, the teacher will "provide clear procedures for determining a percentage mark for the report card for a student who has failed to submit one or more assignments for evaluation on time or at all" (p.44). After spending several pages outlining the possible benefits and disadvantages of punitive measures, the policy reiterated, "The professional judgement of the teacher, acting within the policies and guidelines established by the ministry and board, is critical in determining the strategy that will most benefit student learning" (p.46). Once again we see a reference to the partnership of individual professional judgement and the local administration to determine processes. We will examine how this partnership works in practice, and how the respondents treat late and missing assignments when determining the report card grade as part of Section 3.3.

1.6.5.6 Evaluation Summary

Growing Success's chapter on Evaluation established many research questions. Although it is stated that a report card grade should be determined by a student's most consistent achievement with special consideration for more recent,

and that professional judgement plays a significant role in establishing non-mathematical factors in this calculation, how the actual process works was not fully explained. It also does not expound on how professional judgement and the fundamental principles work together when determining student grades, such as converting levels to percentages. Combined with the open position on how to best address late/missing assignments, there are many questions about how this policy works in practice. Much of the interpretation was left to a school's administration to establish a local policy. The questions asked regarding this section will contribute to the discussion with the participants regarding the relationship between individual interpretation and provincial consistency.

1.6.6 Reporting Student Achievement

The next section addressed the standardized report card used for reporting student achievement. Although it does not address professional judgement per se, since this study concerns final report card grades, it is worth taking a brief look. The report card did not change dramatically with the release of *Growing Success*. Nevertheless, the section did generate some research questions. Most importantly, the issue of student behaviour is repeated.

Most of the information in this section is pragmatic. For example, it included a description of report cards, including the physical dimensions and paper quality (p.49). On a standardized template, teachers state the overall percentage grade, the course median, learning skills score, and comments. The report card is almost identical to the template released as part of the New Curriculum. There is

one distinction worth noticing: the box where the overall average is entered for all current courses was eliminated. In fact, not only is the word "average" no longer on any part of the report card – it did not appear anywhere in *Growing Success*. Also, unlike previous documentation that encourages the use of evaluation software because of its "definite advantages" to "calculate" grades (Ontario, 1999a, p.5), *Growing Success* made no mention of electronic means in order to determine grades. The implications of these omissions will be further explored when discussing evaluation software and calculation methods (see Section 3.2.2).

As with prior report cards, teachers were instructed to record the total number of absences and punctuality issues for the report card (p.55). However, although a Secondary School course is comprised of 110-hours of scheduled classroom instruction, there is no minimum number of hours a student must attend class in order to earn a credit. Again, a student's mark is based upon achievement of the curriculum expectations, not behaviours. If attending class is not an overall or specific expectation, missing class and/or being late are disciplinary matters.

Since absenteeism and tardiness are behaviours, they can be addressed in the learning skills and/or comments. Readers were reminded, "to the extent possible, the evaluation of the learning skills and work habits… should not be considered in the determination… percentage marks for subjects/courses" (p.55). However, how do the respondents feel about students who frequently miss class or are habitually late? Does this behaviour play a factor when they are determining the overall grade? Such questions are answered in Section 3.4.1.

To summarize, the topics in this section of the policy relevant to this study is how the participants react to attendance issue and the external means they use to help determine grades (e.g., with evaluation software). These inquiries add to the list of questions of how the participants used their professional judgement to balance the policy of not allowing behaviours to interfere with assessment of learning. Furthermore, how educators use tools to help with reporting student learning is linked with professional judgement and needs to be examined.

1.6.7 The Remaining Sections of *Growing Success*

Not every chapter of *Growing Success* needs to be explored for the purposes of this study. For instance, the ninth section discusses E-Learning (online instruction) and the tenth talks about Credit Recovery. As this material does not address the primary concern of professional judgement and its role on determining final report card grades, there is no analysis of these chapters. Furthermore, Appendix One of *Growing Success* provides detailed about Ontario's Large Scale Assessments; Appendix Two covered the physical structure of report cards; Appendix Three contained resources that do not apply to this study.

1.6.8 Close Reading of *Growing Success* Summary

Reading through Ontario's assessment policy in regards to professional judgement, generated a list of specific research questions. All these questions revolve around the major question: what is the role of professional judgement in determining final report card grades in Ontario's Secondary Schools? The following charts summarize Ontario's assessment policy as described by *Growing*

Success, and graphically organized the questions raised regarding how

professional judgement works in practice. The charts will clarify what insight the

study hopes to provide and why this information is a contribution to the field of

knowledge. Chart 1.1: Determining Report Card Grades provides an overview of

all the factors that influence the making of an Ontario Secondary School report

card. The Fundamental Principles box has been placed at the bottom to signify

ideals such as being fair, transparent, and equitable as the foundation of Ontario's

assessment policy. At the top, a box labeled Administration embodies the

oversight provided by local guidelines, both to enforce the policy and provide

clarification to educators. Above the Fundamental Principles, the Curriculum

Expectations, also referred to as the Content Standards, are used to build

individual courses. All three of these boxes are shaded the same colour to

represent external policy-related factors on report card grades. Note there are no

references to professional judgement on this chart. As demonstrated in this

chapter, professional judgement accompanies all teacher assessment decisions.

Therefore, the Teacher Box at the centre of the chart represents not only the

classroom teacher but professional judgement itself. One of the many actions by

The Teacher is to take the Overall and Specific Expectations and use them to

create opportunities for students to demonstrate evidence of learning. Since the

only difference between formative and summative assessment is how evidence is

used, a similar colour is used on the left and right side of the charge. On the right

side of the chart, Specific Expectations are used to establish daily lessons, which

allows for assessment as and for learning. The cyclical nature of formative

assessment guides the learning process as the teacher uses this evidence of

learning to reinforce certain specific expectations, identify students who require additional assistance. Meanwhile, students are also encouraged to discover how they learn best, and find ways to foster their own success independently. The teacher also breaks the overall expectations into learning categories for summative

Chart 1.1: Determining Report Card Grades
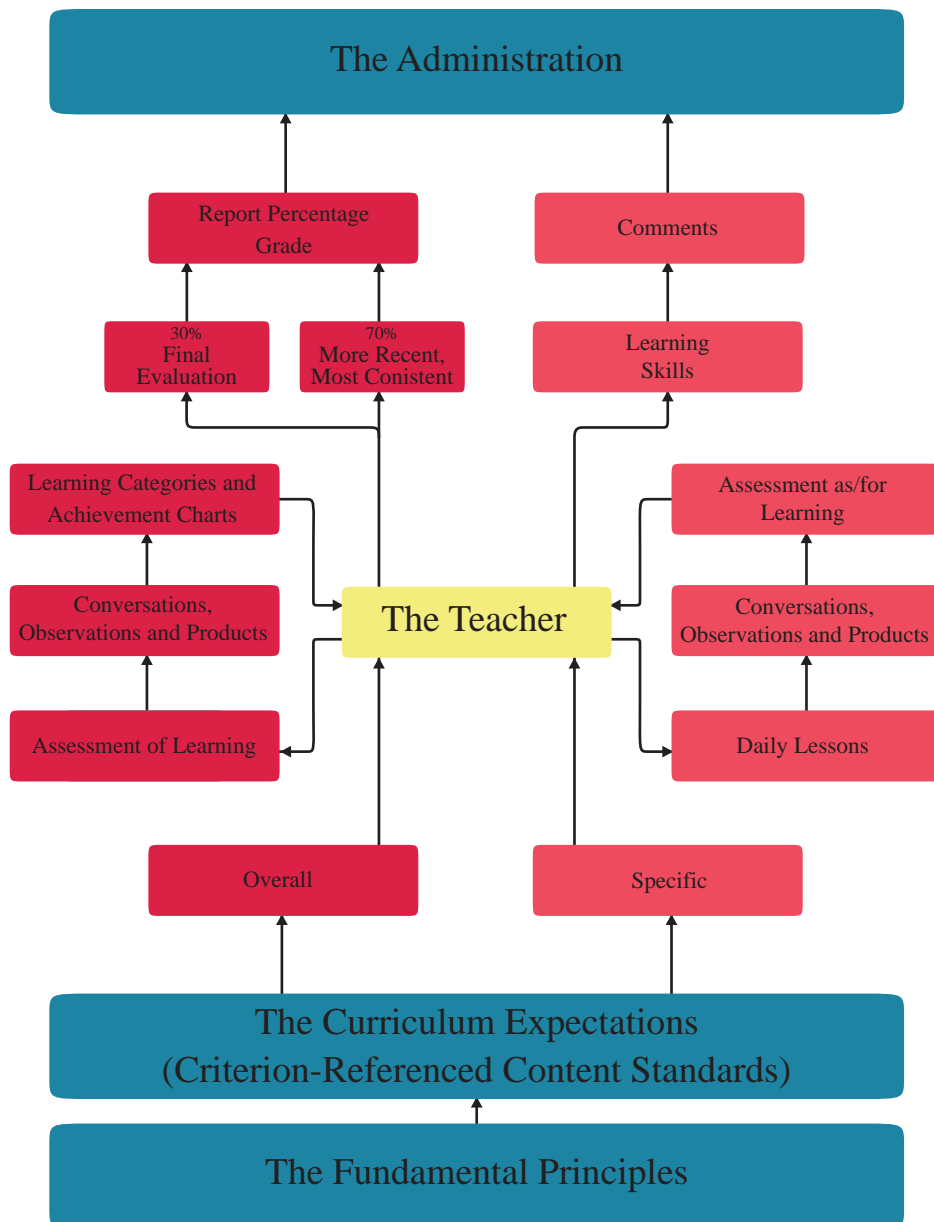
purposes. These summative assessments constitute the course work. Evaluation

of the learning categories is guided by the descriptors on the provincial

achievement chart, so students, parents/guardians, and other stakeholders can have

a transparent look at assessment of learning. The results are recorded by the

teacher, and accumulate throughout the course. Furthermore, since all assessment

is meant to improve learning, a teacher could use the results for formative

purposes by identifying further instruction required to clarify certain expectations.

When it is time to create the final report card, an educator can use any appropriate

evidence to record the learning skills. Meanwhile, evidence of consistent

achievement of the overall expectations, with special consideration for more recent

achievement, is used to determine a student's course work grade. Professional

judgement will assist with pinpointing a percentage grade. Towards the end of the

course, the student will also complete a final evaluation (e.g., an exam), which

tests all or most of the overall expectations. The result of this assessment is then

added to the course work to establish the overall percentage grade. Finally, the

teacher will include anecdotal comments on the student's achievement and

performance. However, the questions raised in this chapter build a case that, in

practice, assessment is much more complex than Chart 1.1 would suggest. Indeed,

anyone familiar with the literature, or has taught any course that requires assigning

a number to the words of students, is challenging. Respondents' thoughts on these

questions will show just how multifaceted assessment in Ontario is.

Chart 1.2: Factors Influencing A Teacher's Professional Judgement

illustrates a framework of professional judgement and its role in determining

report card grades as described in *Growing Success*. It is a close up of "The

Teacher" box found on Chart 1.1. Using the Ministry's definition of the term, a

teacher begins with addressing five key knowledge areas: the curriculum

expectations, context, evidence of learning, methods of instruction and assessment,

and the criteria standards. Furthermore, this knowledge is internalized and

Chart 1.2: Factors Influencing A Teacher's Professional Judgement
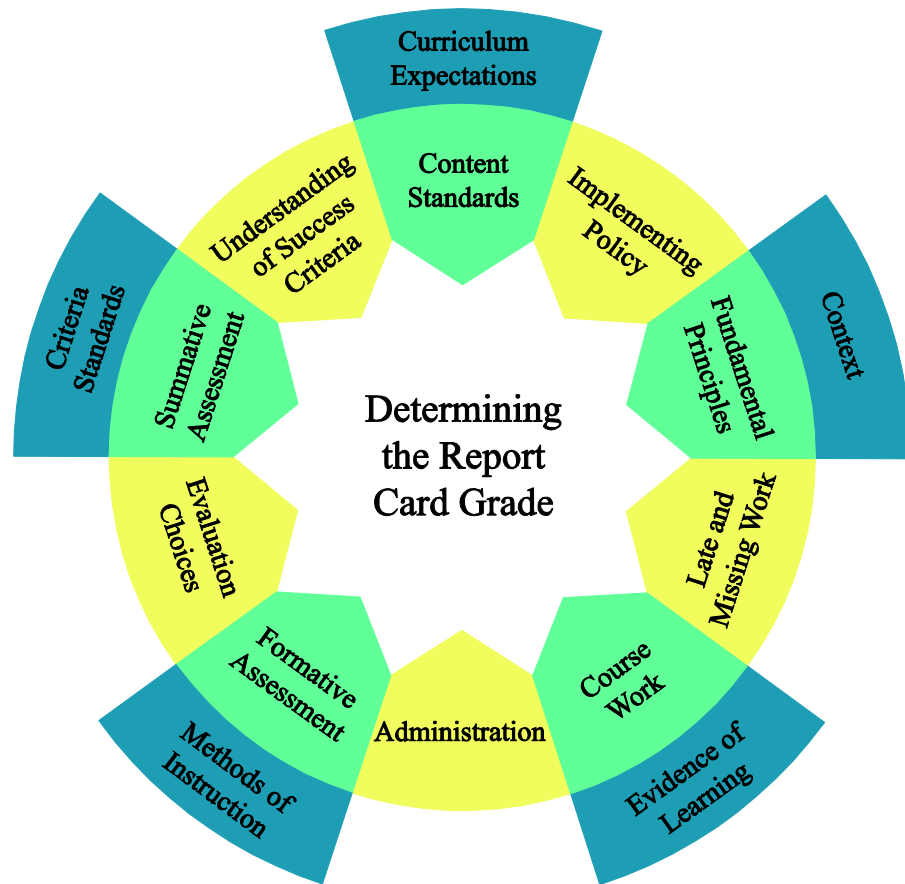
methodically processed, to determine the proper course of action. The definition

also suggested this procedure should become more efficient over time. Chart 1.2

attempts to correspond the five difference types of knowledge with real-world

factors found on Chart 1.1 which need to be internalized: the content standards, fundamental principles, course work, formative assessment, and summative assessment. Also included are other instances where the policy stated professional judgement is necessary: understanding success criteria, implementing policy, addressing late and missing work, following the local administration, and evaluation choices. All of these elements need to be considered in the course of determining a student's report card grade. While Chart 1.2 is a plausible illustration of the way professional judgement is described in *Growing Success*, the actual process and implementation of professional judgement still needs to be explained. One of the goals of this study is to provide an explanation. This is where the research questions come into play. By talking to Ontario Secondary School educators about their practices, it is hoped it can be established how professional judgement juggles different demands, establishes priorities, organizes information, and makes sense out of proper assessment procedures to provide reliable and valid report card grades by using Constructivist Grounded Theory as outlined in Chapter Two (see Section 2.3).

Next, Table 1.2: Assessment and Evaluation Rules for the Final Report Card Grade provides a review of what Ontario teachers, according to *Growing Success,* must and must not consider when recording a student's percentage grade. Table 1.2: General Ontario Guidelines for the Final Report Card attempts to focus on the tangible actions an Ontario teacher needs to take when deciding on the report card grade. In short, a student's percentage grade is 70% course work (balanced by learning category) based upon his or her most consistent

achievement, with special consideration for more recent achievement, of the

Overall Curriculum expectations through student products, conversations, and

observations. Mathematics and professional judgement are required to determine

this level of achievement. Furthermore, the teacher must ensure that the focus is

Table 1.2: General Ontario Evaluation Guidelines for the Final Report Card

Teachers Must	Teachers Must Not
- Base 70% of a student's grade on individual course work (consisting of numerous and varied assignments that show a balance of the learning categories), reflecting the most consistent achievement of a course's Overall Expectations, with special consideration for more recent achievement, using both mathematics and professional judgement - Most consistent/more recent achievement includes student products for evaluation, observations, and conversations - Recognize the difference between formative and summative assessment - Base 30% of a student's grade on a final evaluation that reflects most or all of a course's Overall Expectations administered at the end or near the end of a course - Use Criterion Referencing as per the achievement chart - Ensure assessment reflects the fundamental principles, especially in the interest of being fair, transparent, and equitable - Consider missing work as part of determining a course mark, as long as the result does not misrepresent actual achievement - Follow the guidance of the local administration on all assessment and evaluation issues	- Consider the 30% final evaluation as part of most consistent/more recent achievement - Include group work or self-assessment - Consider Student Behaviour (e.g., participation and attendance) as part of the overall percentage grade - Use Normative Referencing

on evidence gathered under summative purposes. An additional 30% of the final

grade is encompassed in a final evaluation of all or most of the overall curriculum

expectations at the end or near the end of a course. However, the final evaluation

should not count or influence a decision regarding the course work, even as a

reflection on more recent achievement. Both course work and the final evaluation should only be evaluated with achievement in mind. Any reflections on student behaviour as part of the grade will interfere with the reliability and validity. All assessment of learning should be based on criterion referencing using the province's achievement chart. Normative-referencing is to be avoided. At the same time, an educator needs to keep the fundamental principles in mind, and ensure that all assessment and evaluation is fair, transparent, and equitable. Teachers may apply punitive measures to individual assignments, but professional judgement should ensure that such measures do not interfere with expressing a student's true overall achievement. Should there be any need for clarification, a teacher should consult with a member of the local administration.

However, there are many still areas that require clarification. In practice, what instructions to educators receive from the administration? How do these directives influence decision making? When it comes to the average teacher interacting with students, how much flexibility does professional judgement permit? Are the boundaries of assessment inevitably porous despite the specific guidelines? The dozens of questions raised in this chapter show that there are several policy areas that need further investigation to determine their actual influence on professional judgement as part of this study.

Chart 1.3: Research Inquiries Regarding Professional Judgement and Determining Final Report Card Grades summarizes the questions raised in this chapter as bullet points under three themes: Understanding Policy, Student Behaviour, and Administrative Guidelines. Understanding Policy mostly concerns

what teachers should be doing on the report card, as outlined in Table 1.2. How

well did the participants understand these principles, and incorporate them into

their report card procedures? Student Behaviour raised key points of how a

participant can objectively identify an overall level of achievement in the presence

of ubiquitous subjective-factors such as a student's overall classroom performance.

When an educator has shared a learning experience with another person, how can

social factors be accounted for to ensure a reliable and valid evaluation?

Administrative Guidelines highlights how *Growing Success* is interpreted at

individual schools. For instance, how it wants to see policy carried out, such as

the steps that are taken to calculate/determine grades and how this information

informs interaction between the administration and participants. Organizing the

questions by theme was an efficient way to create a semi-structured interview

guide for the researcher and the participants to inform the narrative of the study.

**Chart 1.3: Research Inquiries Regarding Professional Judgement and
Determining Final Report Card Grades**

The Actual Role and Process of Professional
Judgement in Determining Report Card Grades

Participant Insight

Understanding Policy	Student Behaviour	Administrative Guidelines
• The Fundamental Principles • Origin of Understanding • Criterion vs. Norm Referencing • Levels and Learning Categories • Using Formative and Summative Assessment	• Missing Work • Academic Integrity • Attendance/Behaviour • Outside Factors	• Local Interpretation of Policy • Calculating vs. Determining Grades • Converting Levels to Percentage Grades • Coursework and the Final Evaluation • Addressing Punitive Measures

1.7 Professional Judgement in Ontario Secondary Schools Summary

This Introduction provided a general overview of the topic under investigation: what is the role of professional judgement in determining final report card grades in Ontario's Secondary Schools for students of Business, Humanities and Social Science? The detailed reading of the assessment policy *Growing Success* provided the province's position on the importance of professional judgement, but how to apply it is unclear. A comprehensive look at the document bred numerous inquiries in the different ways educators need to apply professional judgement. This study will provide insight into how different assessment factors come together to the real world. In the complex situation of the Ontario schools, the participants ultimately did their best to make sense of assessment matters by simplifying the process. Although the respondents acted in the perceived best interest of the students, and applied practices based on local instructions and personalized reasoning, there are concerns on how informed professional judgement is being practiced.   When professional judgement is only a list of procedures, it turns into Heuristic Assessment. Validated and transparent decisions must form the foundation of informed professional judgement. It is hoped addressing problems discussed in this study, and proposing how to fix them, will make a valuable contribution to the field of knowledge.

**CHAPTER TWO: METHODOLOGY AND METHODS –
CONSTRUCTIVIST GROUNDED THEORY**

2.1 Introduction

The purpose of this chapter is to outline the constructivist methodology and

methods/research procedures used in this study. A well-crafted research project

requires a detailed rationale for why certain actions were taken. First, the

methodological side of the study is provided, with special attention to

epistemological underpinnings. Next, an explanation of the methods and research

procedures, organized in chronological order, is discussed. This chapter will

clarify how I investigated the role of professional judgement in the Ontario

Secondary School final report card evaluation process.

A qualitative approach was used to determine what professional judgement

meant to the participants. The assessment situation in Ontario needed to be fully

explored to examine the actions of the respondents, their interactions, and

relationship with governing structures such as administration and policy. The

research gathered their voices in an effort to co-construct an interpretation.

Constructivist Grounded Theory (CGT), adapted from the work of Kathy Charmaz

(2006), guided the study. CGT applies a constructivist paradigm to Grounded

Theory Methods (GTM), e.g., constant comparison, theoretical sampling, and

creating conceptual categories. Furthermore, special attention to the grounded

theory tradition of pragmatism, such as abductive reasoning and symbolic

interactionism, refined the iterative design. Adele Clarke's (2005) advice on

situational analysis also helped to incorporate the multiple perspectives of

participants. Postmodern epistemological concerns, such as questions regarding creditability of the findings, are also part of the discussion. By showing how an amalgamation of CGT and postmodernism can be applied to educational research, a methodological contribution to the field of knowledge is made.

As discussed in the previous chapter, a close reading of Ontario's assessment and evaluation policy *Growing Success* showed there are three major arenas, or points of contention, concerning evaluative professional judgement: understanding policy, recognizing the difference between achievement and behaviour, and following administrative guidelines. Informed professional judgement, a teacher's knowledge and active thinking process, was emphasized within the policy as the means to ensure grading is reliable and valid. Since professional judgement, however, is open to interpretation, and comes as a result of constant action and interaction between educators and students (i.e., the daily classroom and evaluating assignments), what does applied professional judgement in Ontario tend to look like? How do the participants turn policy into practice and determine a grade? How are matters such as equity and fairness, consistency and transparency addressed? What are the challenges of following administrative guidelines while applying one's own judgement? How do participants monitor potential biases to ensure assessment improves learning? An analysis needed to account for what, how, and why questions associated with the area of inquiry. CGT provided insight into the problematic social world of grading practices in Ontario's Secondary Schools from the point of view of twenty-four participants to inform the analysis in Chapter Three. The research design provided a means to

articulate concerns that this study categorized as Heuristic Assessment, procedures

that lack reliability and validity but are common within the present evaluation

system in Ontario. Fortunately, conditions that cause Heuristic Assessment can be

corrected through further conversation (see Section 5.3).

2.2 Defining Methodology

As a researcher with an insider's view of the area under investigation, I had

an opportunity to better understand what the participants meant when they

explained what they do, how they do it, and why they do it, when it comes to

determining final report card grades. Since the data was comprised of thick,

situated description, a qualitative approach worked well with an interpretive

analysis to attenuate the information. However, I had to ensure my interpretation

was not an act of forcing meaning (Charmaz and Belgrave, 1996). Knowledge is

subjective, but all potential ideas had to be challenged to ensure that they are based

on data – and not merely on personal opinion. There is no universal paradigm for

qualitative research. Therefore, a clearly articulated methodology was needed.

Harrington explained that using a methodology "is to follow a rationale that

justifies one's selection of these particular methods for a given topic of study"

(2005, p.4). Methodology governs the methods, and it is not to be applied

haphazardly. Certain steps need to be taken, and the reasoning behind how the

methods were chosen and carried out must be well defined. Furthermore, asserting

why the steps were believed to be correct reveals the epistemological stance of the

researcher, which also needs to be justified (Charmaz, 1990; Banister, et al., 1994;

Annells, 1997a; 1997b; Creswell and Miller, 2000; Caelli, Ray and Mill, 2003;

Bryman, 2004; Yates, 2004; Silverman, 2005; Rolfe, 2006b; Greckhamer and

Koro-Ljungberg, 2006; Mills, et al., 2007; O'Donoghue, 2007; Mason, 2010;

Lincoln, Lynham, and Guba, 2011; Milliken and Schreiber, 2012).

2.3 Constructivist Grounded Theory: Background and Use in this Study

The methodology used in this study was based on the CGT of Kathy

Charmaz. This choice was not only suitable for the nature of the research

inquiries, but reflected the "worldview" of the researcher (Annells, 1996, p.379;

Clarke, 2007). As the name suggests, it takes a constructivist stance: knowledge

and meaning are social constructions, the researcher and participants co-create an

interpretation of social processes, and the voices of the participants are emphasized

(Guba, 1990; Charmaz, 1995a; 1995b; 2000; 2005; 2006; 2008b; 2009; 2011;

2012; Guba and Lincoln, 2005; Mills, Bonner and Francis, 2006a). However,

CGT is not without its controversies. As a result of its divided history, the use of

GTM generates debate. Two researchers could provide a similar explanation on

how data was collected, coded, organized, and analyzed – but argue that the

opposing researcher was not actually applying "proper" GTM. A synopsis of

grounded theory's complicated story helps to clarify epistemological and

methodological misconceptions and intentions, as well as how GTM was used in

the study. Furthermore, we can see how CGT addresses inherent weaknesses in

original GTM, including insights from researchers such as Adele Clarke who

applied situational analysis to take the methodology further along the postmodern

turn (Clarke, 2003; 2005; 2007). Throughout this chapter, it is shown how GTM

was applied and modified to suit the needs of the study, not just to establish

trustworthiness and philosophical clarification, but to contribute to the methodology itself as well as the field of knowledge regarding assessment and evaluation (Annells 1997a; 1997b).

2.3.1 The Origin of Grounded Theory Methods

In the 1960s, social science research was dominated by quantitative, hypothesis-driven methodology in the positivist paradigm (Denzin and Lincoln, 1994a; 1994b; 2005; 2011; Guba and Lincoln, 1994, 2005; Kennedy and Lingard, 2006; Lincoln, Lynham and Guba, 2011; Charmaz, 2012). The modernist notion that knowledge was an external entity, and could be observed objectively and demonstrated mathematically and/or deductively, was supreme. Qualitative research, having enjoyed earlier success, was essentially relegated to providing addition details to quantitative studies (Robrecht, 1995; Angen, 2000). Two Sociology researchers from the University of California San Francisco challenged this quantitative authority. Anselm Strauss and Barney Glaser qualitatively investigated the interactions between medical staff, patients, and their families. Both Glaser and Strauss were interested in how meaning is socially created, and using the views of the participants to govern analysis and ongoing data collection. They referred to their approach as grounded theory. They published several works, including a book on their methodology, *The Discovery of Grounded Theory* (1967).

In short, instead of trying to prove a hypothesis, grounded theory strategy creates one through the conceptualization of data. Researchers can use their prior

knowledge as a starting point – but the results should be developed based on the point-of-view of the participants. Analysis and data collection are combined in the form of constant comparison. Glaser and Strauss allowed emerging findings to guide additional data collection as they identified potential areas of interest, referred to as theoretical sampling. Voices of the participants are coded, noted, and organized into increasingly abstract conceptual categories. Once the categories are saturated, i.e., new data tends to confirm what is already gathered, a central idea is identified and provides insight on a social phenomenon. Glaser and Strauss were not interested in testing their findings per se, but generating "theory" (i.e., an explanation or interpretation) that may have a practical application (Glaser and Strauss, 1967; Robrecht, 1995; Locke, 1996; Taylor and Bogdan, 1998; Lomborg and Kirkevold, 2003; Boychuck-Duchscher and Morgan, 2004; Heath and Cowley, 2004; Stainton-Rogers, 2006; Suddaby, 2006; Wasserman, Clair and Wilson, 2009). This study attempted to stay true to the original vision of GTM, while also including major developments in research practices since its inception.

2.3.2 The Pragmatist Tradition in Grounded Theory Methods

It was not that Glaser and Strauss had created an entirely new methodology, but they went against popular contemporary thought and gave qualitative research newfound respect (Denzin and Lincoln, 1994a; Covan, 2007; Charmaz, 2008c; 2014). Glaser had recently received his PhD from Colombia University, having worked with noted sociologist Paul Lazersfeld. Lazersfeld is primarily remembered for his quantitative work, but he applied qualitative methods as well (Bryant, 2009). Strauss had already been in academia for

decades. He was a student of Hubert Blumer at the University of Chicago where he learned tenets of pragmatism, i.e., an emphasis on the usefulness of knowledge, including abductive reasoning and symbolic interactionism (Rennie, 1998; 2000; Suddaby, 2006; Strübing, 2007). Awareness of this pragmatic tradition, which was important to both Glaser and Strauss, helps to explain grounded theory methodology and research decisions in this study (Milliken and Schreiber, 2012; Charmaz, 2014).

2.3.2.1 Abductive Reasoning

Glaser and Strauss's grounded theory is usually said to be inductive (Morse, 2001). However, this claim is misleading (Charmaz, 1990; 2006). The reasoning was based on abduction, a major philosophical contribution by pragmatist Charles Pierce (Clarke, 2007; Charmaz, 2010; 2012). Pierce's theory of inference was a "self-correcting" interplay between abductive reasoning and induction (Rennie, 1998; 2000). Pierce wrote, "Abduction seeks a theory. Induction seeks for facts. In abduction the consideration of the facts suggests the hypothesis. In induction the study of the hypothesis suggests the experiments which brings to light the very facts to which the hypothesis had pointed" (cited in Sebeok and Sebeok, 1981). He also described abduction as "the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea" (cited in Suddaby, 2006, p.639). Simply put, abduction is educated guessing. Researchers investigate data and speculated all possible reasons for why certain actions are happening. Insightful questioning serves as deductive testing for an emerging hypothesis, and additional information is

collected until the most reasonable explanation remains. Hence, Piece saw

deduction as tautological, as the researcher mapped out why the hypothesis was

the most reasonable explanation for the phenomenon. With this method of

reasoning in grounded theory research, findings have the potential to serve a

practical application as they have undergone rudimentary testing.

2.3.2.2 Symbolic Interactionism

Another pragmatist scion, symbolic interactionism, was part of grounded

theory strategy from the start (Milliken and Schreiber, 2012). Symbolic

interactionism developed in the early 20th century by George Mead and was

refined and named by Blumer. The perspective has influenced study in many

disciplines, including education and policy analysis (Strauss, 1958; Blumer, 1969;

Annells, 1996; Heath and Cowley, 2004; Charon, 2010; Charmaz, 2012).

Traditional symbolic interactionism states one's interpretation of the meaning of

symbols (i.e., all things found in the world) influences one's social actions.

Blumer explained that how a person interpreted, and reacted to, symbols, evolved

with the accumulation of additional experience through social interaction. Within

this social process, language was perhaps the most powerful symbol (Cutcliffe,

2000; Charmaz, 1990; 2006; 2014; Bogdan and Biklen, 2007; O'Donoghue, 2007).

Grounded theory methodology was influenced by symbolic interactionism because

it was a practical way to explain the complexity of interactions between the

individual and society (Strauss, 1987; 1993; Corbin, 1991; Robrecht, 1995; Clarke,

2003; Mills, et al., 2007; Charmaz, 2008a; 2008c; Bryant, 2009; Pidgeon and

Henwood, 2010; Milliken and Schreiber, 2012; Chamberlain-Salaun, Mills and

Usher, 2013). By focusing on action and interaction within social structures and

processes, it can be seen that people are constantly negotiating meaning. There are

many possible interpretations of a common symbol, including simultaneously

reasonable and unreasonable reactions within a group. Social worlds and

individuals impact one another, creating a perceived external world as an

extension of the internal world. Individuals interpret and incorporate social

meaning as personal knowledge, truth, and perspective. Strauss in particular

considered the presence of multiple truths and perspectives within a social world

(Strauss, 1993; Mills, Bonner and Francis, 2006b; Clarke, 2007).

2.3.3 Grounded Theory Methods and The Postmodern Turn

The work of Glaser and Strauss was more interpretive than many of their

contemporaries (Bryant, 2002; Urquhart, 2002). With its attention to social

constructionism and participants' viewpoints, grounded theory strategy hinted at

the subjective creation of knowledge. However, the methodology was still within

an overarching positivist paradigm: the notions of "discovery" and "emergence"

suggested that there was an external, objective truth that could be observed by a

passive researcher. Although data was based on the views of the participants,

Glaser and Strauss did not quote respondents directly. They hinted that prior

assumptions of the researcher played a role in GTM, but were unclear on how to

account for bias, e.g., how did their opinions impact interpretation and/or how did

they account for multiple individual perspectives? With the absence of an

epistemological discussion in *Discovery*, researchers came to label grounded

theory within the realm of (post)positivism: although it can only be contemplated,

thus potentially producing multiple interpretations, an authoritative truth exists

separately from the objective researcher. Thus a "grounded theory" was an

abstracted but generalized view of a social process (Charmaz, 1990; 2000; 2008b;

Denzin and Lincoln, 1994a; Annells, 1996; Hallberg, 2006; Lomborg and

Kirkevold, 2003; Kennedy and Lingard, 2006; Pidgeon and Henwood, 2010;

Calman, 2011). In addition to the contemporary social scientists who rejected for

its "unscientific" qualitative design, others rejected GTM for being too positivist.

Glaser and Strauss may have started the qualitative revolution, but

grounded theory was rejected by those who wanted to escape any positivist

assumptions (Denzin and Lincoln, 1994b). In what some researchers refer to as

the "paradigm wars" of the 1970s and 1980s, competing methodologies that

emphasized the interpretivist and subjective nature of knowledge and truth

gradually gained acceptance (Denzin and Lincoln, 2011, p.1). For instance, Hesse

wrote in 1980, "The attempt to produce value-neutral social science is increasingly

being abandoned as at best unrealizable, and at worst self-deceptive" (p.247).

More and more researchers were influenced by the radical epistemological shift of

postmodernism.

Postmodernism is a difficult term to define as it can encompass many

different meanings (Delamont and Atkinson, 2010; Lincoln, Lynham and Guba,

2011). Generally speaking, it utterly opposes positivist assumptions and views

associated with modernist thought (Rolfe, 2006a). In short, postmodernism

questions everything. Postmodernism proposes that empiricism and scientific

method must be remodeled, truth and knowledge acknowledged as subjective, the

role of the researcher addressed, and epistemological pursuits scrutinized (Raskin, 2002). To move away from modernist paradigms is sometimes referred to as taking the postmodern turn (Clarke 2003; 2005). As demonstrated throughout this chapter, the postmodern turn has influence the style of GTM that has been used in this study beginning with how it impacted the nuances of GTM's pragmatic foundation (e.g., Section 2.9.7.1.1).

2.3.4 Pragmatism and the Postmodern Turn

Unfortunately, pragmatism was a casualty of the paradigm wars. As reviewed in Section 2.3.2, even though early pragmatists claimed that knowledge was provisional, truth subjective, and both dependent on interpretation, pragmatism was also associated with natural science. The opinion of the majority established the dominant framework (see Section 2.3.1) that could be used for external, objective truth, which led some pragmatists to the positivist realm. Hence, grounded theory's pragmatic foundation was further evidence of positivism. Furthermore, some researchers incorrectly portrayed symbolic interactionism as only addressing micro-level analysis (Clarke, 2007; Charmaz, 2014). However, due to their mutual flexibility, there are strong links between postmodernism and pragmatism. In fact, it could be said that postmodernism germinated from the subjective aspects of pragmatism (Delamont and Atkinson, 2010). Gradually, postmodernists encouraged a revised look at perspectives such as symbolic interactionism (see Section 2.3.2.2) to recognize its interest in larger social structures (Milliken and Schreiber, 2012). However, how far grounded theory strategy has come around the postmodern turn depends on the

epistemological standpoint of the researcher. Again, The intention of the present

study is to use the strengths of pragmatism and postmodernism to aid an

interpretation of the social world of Ontario schools.

2.3.5 Formation of The Constructivist Paradigm

Postmodernism is not, in itself, a paradigm (Clarke, 2003). Its stances and

refusal to be coherently defined make it unstable as a research design. It is a

critical perspective of the research process, which forces the researcher to question

and defend all aspects of thought in the interest of solidifying findings. It is

perhaps mostly associated with critical theory, but even then its precise role is

controversial (Lincoln, Lynham and Guba, 2011; Delamont and Atkinson, 2010).

Since it goes against positivist and post-positivist thought, it aided the

development of the constructivist paradigm. This paradigm would have a major

influence on the development of GTM, and in turn, on this study (Kennedy and

Lingard, 2006).

Constructivists are researchers who are "passionate participant[s] as

facilitators of multi-voice reconstruction" (Guba and Lincoln, 2005, p.196). They

take a strong interpretive stance on data, and emphasize the researcher as a co-

creator of knowledge. Moreover, the voices of the participants, are brought

together to form an agreement on a social phenomenon (Geertz, 1973; Lincoln and

Guba, 1985; Guba, 1996). However, its strength is also its weakness: by focusing

on the subjective nature of knowledge, and acknowledging the presence of

multiple interpretations, technically all constructivist projects could be considered

equally true (Bryant, 2009). Fortunately, by borrowing from postmodernist tenets, one can address this weakness. This study is evidence that by combining grounded theory strategy with the constructivist paradigm and postmodern perspective, convincing research can be produced. However, to understand the formation of CGT, further details from the story of grounded theory are needed.

2.3.6 The Glaser-Strauss Split

The Discovery of Grounded Theory was not a research manual. It was a defense of Glaser and Strauss's procedures to the existing quantitative authority (Bryant, 2009). Again, the book does not specifically discuss epistemology, such as the theoretical aspects of emergence (Charmaz, 1990; Bryant, 2002; Urquhart, 2002; Piantandia, Tananis and Grubs, 2004; Chamberlain-Salaun, Mills and Usher, 2013). After the 1960s, Glaser and Strauss went their separate professional ways. As a result, there were unanswered questions concerning the exact role of extant literature, coding procedures, and prior assumptions in grounded theory strategy. This lack of information had two major consequences: it provided ammunition for those who opposed the methodology as positivist/postpositivist and the spread of misconceptions regarding its application.

Glaser, although leaving academia, continued to publish works regarding grounded theory, such as *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory* (1978). The book helped to clarify some procedures but was still lacking epistemological details. Other researchers tried to glean advice from Glaser, which tended to contribute to poorly-designed research projects with

missing theoretical positions, deficient literature reviews, and faulty inductive reasoning. Nevertheless, the idea of grounded theory methodology gradually grew in popularity among professional and amateur researchers, including doctoral students (Stern, 1980; Charmaz, 2000; Bryant and Charmaz, 2007a; Charmaz, 2008c).

In addition to Glaser's publications, students of Strauss had a chance for direct clarification (Bryant, 2007). However, differences between Glaser and Strauss became more apparent, possibly making their eventual split inevitable (Stern, 1995a; 2009b; Melia, 1996). Strauss received numerous requests for a handbook on grounded theory strategy (Melina, 1996; Bryant, 2009). Strauss responded with *Qualitative Analysis for Social Scientists* (1987), essentially a collection of lecture notes, as well as a paper with his protégé Juliet Corbin (Corbin and Strauss, 1990). However, a more comprehensive overview was still desired. In 1990, Strauss and Corbin released *The Basics of Qualitative Research*. The book provided a series of suggested guidelines for a grounded theory project.

Glaser was livid. He argued that Strauss and Corbin had written a book about forcing meaning, advocating description, and supporting verification – all distortions of grounded theory strategy (Charmaz and Belgrave, 2012). He then published his own *Emergence vs. Forcing: Basics of Grounded Theory Analysis* (1992), where he infamously lambasted his former partner with vitriolic criticism. Glaser polemically highlighted what he saw as errors by Strauss and Corbin, such as using inquiry statements instead of clear research questions, applying an over-

zealous coding procedure, not discussing saturation – and, what he saw as, not crediting his contributions to the methodology.

In the aftermath of these publications, grounded theory strategy was bifurcated between those who chose to follow the "Glaserian" and "Straussian" influence, also referred to as "Classic" and "Evolved," respectfully (Stern, 1995b; Locke, 1996; Morse, 2009). Researchers got what they wanted: advice on how to apply grounded theory. Now with two guides, there was more applicable information, and the methodology continued to gain in popularity. However, the split was so severe, Glaser made it clear that anyone following Strauss and Corbin could not claim to be using grounded theory strategy, but "Full Conceptual Description" (Glaser, 1992, p.3). Grounded theorist Stern declared that the effectiveness of the methodology had been eroded (1995a). This debate hurt its development (Corbin, 1998; Greckhamer and Koro-Ljungberg, 2006).

2.3.6.1 Eroded Grounded Theory Methods

The popularity of *Basics* was a mixed blessing. On one hand, more attention was paid to grounded theory strategy. On the other, it misconstrued the methodology (Bryant, 2007; 2009). Glaser's publications also added to the interest, and his parsimonious version of grounded theory still gathers a significant following (Stern, 2009a). However, conflict between the two sides, and overall criticism from third parties, caused confusion over what is "correct" and "incorrect" GTM. Frankly, both approaches were flawed due to their continued commitment to an overall positivist paradigm and prescribed methods (Charmaz,

2008c; 2010; Lincoln, Lynham and Guba, 2011). They both suggested the presence of an external truth that could be generalized, while struggling to explain theoretical underpinnings and what the research actually accomplished (Charmaz, 2000; Lomborg and Kirkevold, 2003; Greckhamer and Koro-Ljungberg, 2006).

2.3.6.2 Similarities Between Glaserian and Straussian Approaches

As both Glaser and Strauss claimed to continue the discussion started in *Discovery*, similarities can be noted (Melia, 1996). For instance, they both converted from a two-step to a three-step coding system in order to form conceptual categories. Through constant comparison and theoretical sensitivity, a core category is established to explain the area under investigation. The focus was still placed on participant perception. However, both sides stepped away from GTM's pragmatic roots, suggesting a more inductive, as opposed to abductive, base of reasoning (see Section 2.3.2.1). Procedures were also appeared formulaic (Charmaz, 2010). Furthermore, the importance of social constructionism was lost. Without a sufficient theoretical discussion of how conclusions were reached, both approaches only offered generalized findings with positivist leanings (Bryant, 2002; Kennedy and Lingard, 2006; Charmaz, 2008c).

2.3.6.3 Differences Between Glaserian and Straussian Approaches

Although they both have similar procedures, when Glaserian and Straussian approaches are directly compared, they become two different methodologies (Charmaz, 2000). Researchers who think they can borrow and mix whatever they prefer must do so with extreme caution (Locke, 1996; Boychuck Duchscher and Morgan, 2004; Heath and Cowley, 2004). For example, regarding the role of the researcher, Glaser took a strong positivist stand insisting that the

researcher remain a neutral party. Strauss and Corbin (1990) were more interpretivist in this regard, stating that researchers need to be aware of imputing their own beliefs on data, i.e., acknowledge the ubiquitous researcher. However, Glaser (1992) claimed the Straussian approach granted permission to force meaning. Glaser also criticized Strauss and Corbin's claim that literature could be used to help form research questions. To Glaser, a review of the literature is to be avoided because it can lead to hypothesis testing instead of hypothesis generation. Moreover, a researcher does not even need specific research questions, but only "abstract wonderment" (Glaser, 1992, p.22).

The most significant difference between Glaserian and Straussian approaches is the nature of categories. Although Glaser never claimed GTM was easy, he attempts to make it easier to understand. He advises researchers to constantly compare all objective data and allow theoretical sensitivity to guide the establishment of categories until saturation is reached, at which time a core category emerges. In his words, "Categories emerge upon comparison and properties emerge upon more comparison. And, that is all there is to it" (1992, p.43). Emergence is portrayed as its own form of verification that could advance knowledge (Rennie, 1998). In other words, Glaserian/Classic Grounded Theorists claim to discover a theory that provides an explanation or prediction regarding a social process – but there is no need to test the theory. However, according to Miller and Frederick (1999), one cannot claim to have a theory unless it can be tested. Glaser (1999) stated researchers were free to follow-up with a separate study if verification, not theory, is the ultimate goal. Nevertheless, suggesting that the core category emerges from data continued the major criticism of GTM: that

there is an external truth that can be generalized by a neutral observer (Charmaz, 2008c). Epistemologically, what Glaser means by emergence, theory, and data, is unclear (Lomborg and Kirkevold, 2003). In regards to how the data is analyzed, categories are actually based on eighteen pre-existing coding families, which require background knowledge in Sociology to properly work with (Keller, 2007, p.200). Also, in Glaser's approach, there is no room for situated description, only abstraction (Glaser, 1998). His dictum "all is data" appears to be contradicted because he does not include how the researcher influences the findings or the research context because, in his view, it has nothing to do with abstraction. Clarke wrote, "Neither history nor geography nor culture, much less gender, race, class, or ethnicity, necessarily matter in the Glaserian world" (2007, p.431). Consequently, he disregards the wide range of evidence that abductive reasoning is meant to process to help the hypothesis to explain the facts. Much more detail could be provided on why this study did not follow Glaser's research methods, but citing major philosophical differences alone should suffice. Glaser's rejection of situated description – and his adherence to positivism – does not allow for a proper analysis of the evaluation process in Ontario.

In regards to the Straussian approach, in its ambition to explain GTM, *Basics* was difficult to follow. Its coding paradigm and conditional matrix for forming categories, although broken down into multiple steps, were convoluted (Bryant, 2002; 2007; 2009; Suddaby, 2006). It was also suggested the core category is made to fit the categories, confusing what was meant by constant comparison. There was also discussion of using personal observations to help explain social processes, hinting that the researcher was forcing meaning on the

findings. Although *Basics* was more open to situated description, supported

pragmatism, and recognized the role of the researcher, it also indicated the need

for an objective observer. As a result, the original Strauss and Corbin position

could not be used for this study on the grounds that it was ultimately positivist and

the coding process unnecessarily complicated.

I admit that aspects of Glaser's framework can sound postmodern with

statements such as "all is data" but he inevitably retreats to positivist grounds.

Conversely, Strauss and Corbin can sound rigid, but actually build a postpositivist

structure in which researcher imagination is encouraged. For example, they

encouraged some literature review to help form negotiable inquiry statements,

reflecting the reality that a researcher needs to begin somewhere. As a trained

academic, even Glaser recognized that a researcher could not escape all prior

assumptions. Both Glaser and Strauss/Corbin were indicating that the focus

needed to be on data, and research questions were also subject to theoretical

sensitivity, but explained these positions in different ways.

Regardless of concerns about Glaserian or Straussian GTM, both continued

to be used for research projects in numerous fields (Clarke, 2009). This popularity

may have contributed to the perception that GTM is easy to apply, due to

misconceptions about not needing a literature review, about letting the participants

do all the talking, and about there being no need to provide an epistemological

position (Suddaby, 2006). Lack of researcher theoretical knowledge lead some to

believe that GTM is interpretive because of the emphasis on the views of the

participants (Bryant, 2009). Such studies tend to only cite *Discovery*, along with

Basics and/or *Emergence*, and after a light discussion of subject literature and

material from the participants, with few-to-no epistemological points, claim to have produced a grounded theory – when they have provided only qualitative description (Lomborg and Kirkevold, 2003; Kennedy and Lingard, 2006). According to Charmaz, "For postmodernists, grounded theory epitomized distanced inquiry by the objective experts who assumed their training licensed them to define and represent research participants" (2008b, p.400). Fortunately, antecedently to the chronological split of Glaser and Strauss, some worked on the social constructionism aspects of GTM. With a constructivist paradigm and/or postmodern perspectives, these researchers reconnected with GTM's pragmatist roots and a stronger epistemological foundation was developed.

2.3.7 Charmaz and Constructivist Grounded Theory

Kathy Charmaz was a student of both Glaser and Strauss and has spent her career working with GTM. Charmaz credited Glaser's influence in her earlier work but gradually developed a model based more on Strauss's ideas (Charmaz, 1983; 1990; 2008c). With her insider knowledge, she could read deeper into texts such as *Discovery*, *Basics*, and *Emergence*. In short, she claimed researchers like her never saw original GTM as objectivist, despite the way it was portrayed by Glaser and Strauss/Corbin (Charmaz, 2008b). For instance, she felt there was a difference between being objective and being fair to the participants (2012). She did not see the focus on action, social constructionism and pragmatic thought (i.e., abduction and symbolic interactionism) she had learned from Strauss, such as his work on social worlds, arenas, and negotiations. Therefore, she petitioned for a constructivist approach to GTM, with a well articulated, transparent epistemology and narrative-style write up to defend findings (1995a; 1995b; 2000; 2001; 2006;

2012; 2014). Her work was approved by other researchers, such as Anthony

Bryant, who had initially rejected GTM because of its perceived positivist position

(Bryant, 2002; 2003; 2007; 2009). She did much to effectively modify GTM and

further increased its popularity (Mills, Bonner and Francis, 2006a; Pidgeon and

Henwood, 2010).

2.3.7.1 The Role of the Researcher

Constructivist grounded theory takes a much different stance than the

Glaserian and Straussian approaches, especially regarding the role of the

researcher. When interpreting a particular social process, the researcher

recognizes that s/he builds an interpretation with the participants: analysis, data,

and the researcher, become one (Bryant and Charmaz, 2007a). The researcher

understands that knowledge is situated and subjective in the first place, and the

result of data interaction is only one explanation of any number of possible truths.

Results offer insight, not a generalized theory. To avoid accusations of findings

based on mere opinion, or in my case, forcing my views on a topic I am close to,

the researcher needs to efficiently re-construct for the reader how the project was

put together to explain the reasoning behind the interpretation (see Section

2.9.7.1.2). The reader will then decide the merits of the conclusion.

Consequently, the reader joins the researcher and participants in the co-

construction of meaning (Clarke, 2003, 2005, 2006, 2009; Charmaz 2006, 2007,

2008b; 2009; Bryant and Charmaz, 2007b; Mills, Bonner, and Francis, 2006;

Clarke and Friese, 2007; Denzin and Lincoln, 2010; Pidgeon and Henwood, 2010;

Charmaz and Belgrave, 2012).

2.3.7.2 Importance of Abduction

By examining interaction, researchers can speculate about connections between the "whats" and "hows" of the phenomenon in order to determine the "why" (Gurbrium and Holstein, 1997, p.200; Charmaz, 2012). New questions are created and become part of a cyclical progression of comparison and conceptualization in the spirit of abduction. "Abductive reasoning resides at the core of grounded theory logic," wrote Bryant and Charmaz; "it links empirical observation with imaginative interpretation, but does so by seeking theoretical accountability through returning to the empirical world (2007b, p.46). Hence, the ideas that emerged from grounded theory strategy are the result of complex reasoning grounded in data. More clarification was provided by Reichertz, who said, "Abduction is therefore a cerebral process, an intellectual act, a mental leap, that brings together things which one had never associated with one another: A cognitive leap of discovery" (2007, p.220). No one can know everything about a situation, but the more informed researcher is in a position to make better connections and see what is not always obvious (see Section 2.9.7.1). Constructivists do not expect a core category to emerge, but intend to create one thanks to their own interaction with rich, descriptive data representing a diverse range of respondent's views (see Section 2.13.1).

2.3.7.3 Influence of Symbolic Interactionism on CGT

Pragmatism also influenced CGT in regards to symbolic interactionism (see Sections 2.10.3.2 and 2.10.5). GTM is in itself a theory/methods package

congruent with the goals of symbolic interactionism (Clarke, 2005). Charmaz

(2000; 2014) explained that the research product is an image of the social world,

capturing meaning as explained by the participants. Symbolic interactionism

allows a researcher to be "theoretically agnostic," and approach data looking for

possibilities (Charmaz, 2012). The constructivist creates knowledge with the

participants, and part of the construction is ascertaining how people assign

meaning to a perceived external world (see Section 2.9.7.1.1). With participants, I

could explore how they see themselves within a process, how the process shapes

their self-image, and how the process establishes concepts. People do not simply

act and react, but they also reflect and anticipate. They demonstrate their

understanding of actions, including communication, through their own actions.

Charmaz (1980) also emphasized the importance of shared language and social

processes as part of the negotiation of meaning, especially when people encounter

morass.

2.3.7.4 Flexible Procedures

Another aspect that distinguishes CGT from Classic or Evolved GTM is

flexible procedures (Charmaz, 2000). There are epistemological restraints, but the

proposed criteria are more elastic (see Section 2.9). There are no coding families

or matrixes to organize data, only the sensitivity of the researcher to code and sort

the voices of the participants. That said, there is similar terminology with

Glaserian and Straussian GTM, e.g., coding, constant comparison, theoretical

sampling, etc. Constructivists also seek to gather data until categories are

sufficiently saturated by focusing on processes, as opposed to topics or themes,

because action better demonstrates connections in the data (Charmaz, 2006; 2012; 2014). Like the rest of the methodology, the literature process is flexible (see Section 2.13.2). Generally speaking, the researcher can decide the circumstances of engaging in a thorough literature review. For instance, it may be needed to help with forming research questions, or delayed until after the analysis if the research wishes to guard against adding to prior assumptions (McGhee, Marland and Atkinson, 2007). The present study needed to consult Ontario policy to form research statements, while reserving the subject literature until after data gathering.

2.3.7.5 Reflexivity

The postmodern turn can also be noted in CGT (Morse, 2007; Willig, 2008; Charmaz, 2009; Calman, 2010). Instead of taking anything as a given, the constructivist researcher reflects deeply about the situation (Kinchelo and McLaren, 2007). This introspection includes any possible conditions, including what the researcher personally brings into the analysis, that have consequences on the social process under investigation (Bowers and Schatzman, 2009). It is also important to explicate these points in the final report. Furthermore, all terminology must be explained because "without epistemological connections, grounded theory is not grounded anywhere or, alternatively, it is grounded everywhere. In this case, it becomes an empty method or a text without contextualized meaning" (Greckhamer and Koro-Ljungberg, 2006, p.746). For example, the opening chapter of this study noted my own personal paradigm, some facts about Ontario and its school system, and a synopsis of *Growing Success –*

because they are all conditions impacting the interpretation of how professional

judgement is practiced (see Section 1.3).

2.3.7.6 Glaser's Criticism of Constructivist Grounded Theory

Glaser holds a firm positivist position. Similar to his attacks on Straussian

GTM, he called CGT a "misnomer," identifying it as mere "Qualitative Data

Analysis" (2002). Constructivist data, information that is created through

researcher interaction, in Glaser's opinion, has a minute influence on grounded

theory. The datum is there to be observed; the researcher works with it as it is and

not by what it is envisioned to be. Glaser wrote, the researcher's job is not for

"accurate description," but "transcending abstraction" (2002). From his positivist

view, what concerned qualitative interpretivists like constructivists – did not

concern him. Granted, respondents have multiple perspectives. However, with

constant comparison, researcher predilection is neutralized, and the common idea

beneath these perspectives emerge as categories (see Section 2.11.2). There is no

need to consider situated description if the researcher has saturated the categories

and allowed the key discovery to emerge. Again claiming ownership of the

methodology, he instructed those who are focused on "story telling" not to refer to

their methodology as grounded theory (2002).

2.3.7.7 Merger with Straussian Grounded Theory Methods

As Charmaz developed her constructivist approach, Straussian GTM

continued to evolve. Just as Strauss and Corbin did not engage Glaser in a public

debate, they did not argue against the constructivist view. Instead, a series of

publications (1994; 1996; 1997) elucidated the position of Strauss and Corbin. They defended *Basics* as a supplemental text for novice GTM researchers. Information on epistemology could be obtained elsewhere. In 1993, Strauss also published *Continual Permutations of Action* where he reconnected with symbolic interactionism (see Section 2.4.1). Before he had a chance to elaborate on of the pragmatic role of GTM, he died in 1996. A second edition of *Basics* was released in 1998 with numerous revisions, including a conditional/consequence matrix and elaboration of the axial coding process to examine the relationship between categories. The edition claimed it was "not a recipe book" (p.ix). Nevertheless, it was still seen as positivist by researchers such as Charmaz (2008b). Corbin (1998) continued the defense of their work, stating *Basics* was an attempt "to put into words what is a very difficult process to convey: Anselm Strauss's way of thinking and working with data" (p.121). However, researchers such as Mills, Bonner and Francis (2006b) have observed a more interpretivist leaning in the second edition of *Basics*, such as, "Theorizing is the act of construction" (Strauss and Corbin, 1998, p.25). In 2008, Corbin released a third edition of *Basics*, with further interpretivist revisions. Statements like, "when we share a common culture with our research participants… It makes sense, then, to draw upon those experiences to obtain insight into what our participants are describing," sounded like a constructivist approach (Corbin and Strauss, 2008, p.80). Furthermore, Charmaz (2012) stated the third edition was "much closer to what I've been using." Corbin (2009) confirmed that she, like Charmaz, was focusing on social constructionism. Therefore, one could either state that CGT is a third school of GTM thought, or it has absorbed the Straussian approach. Corbin felt,

"I personally don't see the purpose in all this hoopla about method. One could argue and discuss methods all day. In the end, it doesn't matter. People will choose the method that most speaks to them and they will use it in ways that make sense to them." (2009, p.52)

2.3.7.8 Counter-Argument to Glaser's Criticism

Still, Glaser's concerns about CGT must be addressed. He was primarily focused on protecting the label of "grounded theory" similar to a copyright. Glaser based his opposition on claiming any constructivist issues are qualitative at large, and have nothing to do with GTM. When constructivist concerns were applied to GTM, he applied circular reasoning in that GTM had been misunderstood because it has almost nothing to do with constructivist thought. To address this dilemma, it is useful to remember the difference between methodology and methods. Methodology is the thinking aspect behind procedures as represented by the methods. How the methods are applied depends on the researcher's methodology, which includes an epistemological position. Glaser did not explain his epistemological stance (Greckhamer and Koro-Ljungberg, 2006). To suggest Charmaz was applying methods incorrectly only because she was taking a constructivist stance, does not clarify his position. Furthermore, Charmaz did not dismiss Classic GTM, but suggested that constructivist ideas could address many of its criticisms (Bryant, 2003). A sign of positivist thought is a right or wrong mentality, while the postmodern position asserts that all sides deserve to be heard (Rolfe, 2006a). Therefore, by the tenets of constructivism, Glaser's views must be acknowledged as one of many possible truths (Bryant, 2009). That said, this study disagreed with the epistemological assumptions of Glaser and his claim

that only a certain set of research procedures referred to as GTM. Moreover, to

argue over ownership of a research term is not a good use of resources (Strauss

and Corbin, 1994; Melia, 1996; Dey, 1999). This study used an interpretation of

what is believed to be good use of GTM within a constructivist model, and

recognizes CGT as a legitimate methodology.

2.3.8 Why Use Constructivist Grounded Theory?

This study used CGT because, given the nature of the research questions, it

was the ideal methodology. Choosing the most appropriate methodology to

answer certain questions is the mandate of all researchers (Punch, 2009). I was

investigating an area where little prior research had been done (Kennedy and

Lingard, 2006). The goal was to get into the phenomenon as much as possible

(Charmaz, 2012). Also, in the social process under investigation, situated

description is vital to decision making. Even the definition of professional

judgement advised teachers to consider "context" as part of evaluation (Ontario,

2010a, p.152). Therefore, discussing situated description, and including it in the

analysis, was a necessary part of the study (see Section 2.4.1). Furthermore, to

treat data as an external entity diminishes the human aspect of research. Not to

consider and/or confirm what respondents meant by their words is to look upon

them as data-generators instead of people. Datum is not external; it only has

meaning as a result of researcher-participant interaction, just as all meaning is the

result of one type of interaction or another (Charmaz, 2004). In regards to

accusations that the constructivists impose their own values on the data, Charmaz

countered, "claims of value-free neutrality assume, paradoxically, a value

position" (Charmaz, 2006, p.132). Just as it is discussed in original grounded theory, the constructivist can use background knowledge to his or her advantage, as long as it does not contribute to a hypothesis-driven analysis (Clarke, 2005; Dey, 2007; Gibson, 2007; Mruck and Günter, 2007; Pidgeon and Henwood, 2010). Again, the latent key to solving most research concerns is awareness of what steps are being taken, and why, as well as potential pitfalls. As Denzin stated, "Grounded theory's ground, and the spaces it encompasses, are always constructed, never bedrock solid, always nuanced, and potentially dangerous. The ground itself is a function of the researcher's shifting relationship with the world" (2007, p.458). Hence, one must not take a cavalier attitude to a research project. Furthermore, no matter how fastidious the constructivist researcher, there will always be concerns that the findings are only an interpretation and do not advance knowledge. All the researcher can ever do is not be overly concerned with winning over ever reader, but to be credible by doing one's best to have every reader leave the study feeling that the explanation is based on sufficient data and is a reasonable construction of meaning (Creswell, 2008; 2012).

2.4 Clarke and Situational Analysis

Charmaz once described CGT as "a middle ground between postmodernism and positivism" (2000, p.510). Such statements reinforce the positivist origins of GTM (Glaser, 2002; Greckhamer and Koro-Ljungberg, 2006). Adele Clarke, another former student of Strauss, desired to sever this connection. She termed her project Situational Analysis (SA), as it took a critical look at how everything surrounding the area of the interest is fused as a single situation. Therefore, one cannot analyze a process without carefully looking at the social

world in which it exists. Charmaz has approved blending postmodern perspectives with the pragmatic tradition of CGT in order to move further around the postmodern turn. For example, Clarke instructed researchers to use a range of variance to help organize situated data, along with analytical diagrams, into conceptual categories. Although this study focused more on interactions than making the situation the centre of attention, SA was a tremendous help with my overall analysis.

2.4.1 Understanding Social Worlds with Situational Analysis

As CGT gained acceptance, Clarke (2003; 2005; 2007; 2009) wrote about infusing GTM with the postmodern perspective in order to take the methodology along the postmodern turn. She believed Strauss, and his work with symbolic interactionism, was making this turn because he looked at "conceptualization of social worlds and arenas as modes of understanding the deeply situated yet always also fluid organizational elements of negotiations" (Clarke, 2003, p.556). Whereas positivists like Glaser tried to apply parsimony by disregarding the situation and its infinite description, many postmodern researchers embraced the chaos in order to be closer to the world they are examining (Parkhe, 1993). As Miles, Huberman, and Saldaňa wrote, "The goal is not to impose some artificial order onto the messiness of everyday living and working but to bring enhanced cognitive clarity for the analyst's interpretations of the people he or she is learning from and about" (2014, p.180). To comprehend what the participants are saying, the researcher needs to be aware of their milieu (Hall and McGuinty, 2002). However, there is still the matter of organizing complex information into something ostensive. For example, it is not unusual for CGT projects to identify negative cases for data that

do not fit the overall pattern. Negative cases are an example of the positivist legacy. If data were from the situation, and the research intended to represent the situation, then they must fit into the conceptual design without forcing. Clarke explained, "if one seeks to understand a particular social world, one must understand all the arenas in which that world participates and the other worlds in those arenas are the related discourses, as these are all mutually influential/constitutive of that world" (2005, p.48). Clarke's SA, treating GTM as a theory/methods package, provided advice on how to work with arduous data. SA involves complex diagramming that she called situational, social worlds/arenas, and positional maps (Clarke and Friese, 2007). To summarize, the researcher reflects on the situated data, including human and non-human factors, and their relationship to the overall discourse in order to make connections. The end goal is not necessarily to provide diagrams for a study, although they can prove useful (see Section 2.9.6). Instead, they can be used to better see how data connects to express conceptualization, such as a range of variance – a way of constructing confluent categories that include all data (see Section 2.12.2).

2.4.2 Applying Situational Analysis

Clarke put the situation, instead of the actions and interactions, at the centre of analysis (Milliken and Schreiber, 2012). Therefore, her approach influenced the study less than Charmaz, but the postmodern ideas were certainly helpful. Schools are a social world, and the matters surrounding professional judgements are arenas, which "are usually sites of contestation and controversy. As such, they are especially good for analyzing heterogeneous perspectives or

positions and for analyzing power in action" (Clarke, 2009, p.199). Furthermore,

Clarke underlined the importance of using symbolic interactionism to analyze the

situation (see Section 2.10.3.2). Working within the structure of policy, but also

with other people, brings together myriad factors that need to be negotiated. By

considering how the situation itself influences professional judgement as a social

process raised many questions not only about why participants evaluate in a

particular way, but about the reliability and validity of grades in Ontario. *Growing*

Success described professional judgement as the ability to weigh different factors

when determining a student's achievement and not allowing non-achievement

factors to interfere with grading. However, if all actions are part of the situation,

how does professional judgement determine what should count towards a final

grade? Using a range of variance to discuss such matters was invaluable because

it "analytically allow[ed] the possibility of multiple major processes and that some

processes may even be paradoxical or contradictory" (Clarke, 2005, p.16).

It is not the intention of SA to establish a new school of thought of GTM.

SA is "analytical tools that can be used on their own with discourse data and/or

along with and complementing other theoretical and analytical approaches"

(Clarke, 2005, p.146). Charmaz has supported SA as a means to look at processes

within the larger social (Charmaz, 2008b; 2009). The two agree on many aspects

of GTM (Mills, et al., 2007). Regarding Charmaz's constructivist approach,

Clarke wrote that it was the correct direction for GTM, and she hoped to "use

grounded theory methods as flexible, heuristic strategies… My goal is to further

enable, sustain, and enhance such shifts" (Clarke, 2003, p.559). Perhaps the best

example of Clarke's intended direction is her concern about the voice of participants. In the interest of presenting a flowing narrative, some grounded theorists are anxious to fit the data into discernible categories – which can lead to misconstruing the voice of the participants (Gibson, 2007). To avoid voice-related problems, "I propose," Clarke stated, "that we complicate our stories, represent not only the difference(s) but even contradictions and incoherencies in the data, note other possible readings, and at least note some of our anxieties and omissions" (2005, p.15). This approach to this project helped me articulate my interpretation (see Section 2.10.4).

2.5 Prior Assumptions

It is important for a grounded theory project to outline the researcher's prior assumptions (Outhwaite, 2005). Prior assumptions are separate from general concerns about GTM because they are always present, but when accounted for, they need not be obstacles. They are also distinctly different from hypotheses. Certain biases as part of one's experience are inescapable, but can be nullified by acknowledging that nothing is apodictic (Heath and Cowley, 2004). Similar to other grounded theory researchers, Charmaz stated that "believing that researchers can remain uncontaminated by prior theories and research literatures is epistemologically naïve and rather silly" (2008b, p.135). Assumptions are always present at the start of a research project, as no one can go into a research project as a *tabula rasa*, grounded theory or otherwise (Glaser and Strauss, 1967; Charmaz, 1990; Miles and Huberman, 1994; Urquhart, 2002; Clarke, 2005; Mills, Bonner and Francis, 2006a; Clarke and Friese, 2007; Strübing, 2007; Bryant, 2009;

Wasserman, Clair and Wilson, 2009; Calman, 2011; Charmaz, 2012). A researcher must have some prior knowledge to recognize that GTM is an appropriate approach for a certain area of research. As Dey has stated numerous times, "an open mind does not imply an empty head" (1999, p.251; 2007, p.176). The goal is to not allow the presence of presumptions cloud the analysis, particularly when the researcher is an insider (Willig, 2008; Clough and Nutbrown, 2012). Instead, confronting assumptions can be the basis for creativity and recognizing the ostensible (Cutcliffe, 2000).

At the start of the study, I thought professional judgement was merely a façon de parler for positive mark adjustments. This assumption included whether or not to apply punitive measures, which were forbidden by the OME. This presupposition was mostly based on my first year of teaching when I was unaware of many assessment and evaluation issues. For instance, it was assumed that visible student effort influences evaluation, i.e., hard working students were more likely to receive a propitious grade adjustment. Likewise, students who did not hand in work on time were likely to receive an implicit punitive measure. Consequently, I can see where the research was initially approached with questions surrounding how and why teachers decide to whether or not to adjust marks, and what was the range of these adjustments.

Although professional judgement as a practice does involve adjusting computer-generated grades, there is much more to it. For instance, the degree to which local administrations control professional judgements as part of shared and shadowed practices was unexpected. I was surprised to learn, for example, that an

85

Ontario educator is to separate course work from the final evaluation when determining a report card grade. Despite specific instructions from the OME, local administrations usually ignore this rule and encourage teachers to use the final evaluation to help determine adjustments when converting levels to percentages. We will discuss shared and shadowed practices in detail throughout the study.

It can be honestly said that data were analyzed with an open mind in order to make connections. If conclusions were similar to certain prior assumptions, it can simply be the result of knowledge in the subject area and not the desire to prove hypotheses. All points made in the study are supported by the voices of the participants. Nevertheless, honest questions needed to be asked of whether or not evidence supported a conclusion, or if there was forcing meaning. A genuine feeling of surprise and/or reassurance by the data usually indicated the former. As Angen (2000) wrote,

> "Our values and beliefs will show themselves in our actions whether we stop to think about them or not. We do not live in the world as if nothing mattered, as if everything was relative; rather we live in constant meaningful interaction with people and things, continually, if not consciously, making practical and ethical choices about how to act and interact." (p.384-385)

The key is to carefully look at all research decisions, and reflect on why the decision was made, and what and how it impacts the project.

2.6 Constructing A Grounded Theory Methodology Summary

One grounded theorist wrote that "every time grounded theory is used, it

requires adaptation in particular way as demanded by the research question, situation, or participants for whom the research is being conducted" (Morse, 2009, p.14). With statements such as this, it is not surprising that some do not see grounded theory as a methodology, but only as a set of methods modified to suit a particular task (Greckhamer and Koro-Ljungberg, 2006). However, when one considers larger developments in qualitative research over the last fifty years, CGT is not alone in its needs to clarify it epistemological its stance for each project. On the bright side, every good CGT project is a methodological contribution to the field of knowledge because of the need to explain its theoretical position and how this position influenced the methods. A critique of the methodology can strengthen it and add to the conversation on good research (Bryant, 2002; Clarke, 2005; Charmaz, 2006; Bryant and Charmaz, 2007a; 2007b; Wasserman, Clair and Wilson, 2009). In the next section, we will look at how the previously discussed ideas influenced the research process, such as how participants were identified, data gathered, and how the voices were used to reify an interpretation.

2.7 Grounded Theory Methods: The Research Procedures

The research procedures discussed in this section are framed by the previously discussed methodology. First, there is a talk on the meaning of qualitative research. Next, sections on Methods/Research Procedures have been divided into two major parts, each inspired by chronological order. There are many cyclical aspects of research; they start at the beginning, and are ongoing throughout the process. Other actions have their time and place, and can be considered as stages. Therefore, we will first look at the ongoing actions, referred

to as Constant Reflexivity, then break down the different research stages. It must be emphasized that all parts of the research project oscillate. However, in order to provide a narrative of the research process, sections need to be discussed one by one, even if they overlap with one or more other sections. Specific examples from the study, along with a series of tables, have been provided.

As previously mentioned, potential respondents were approached with inchoate research questions (see Section 1.4). A basic description of the project was provided and semi-structured interviews were scheduled. Twenty-four Ontario teachers provided their views on policy and their assessment practices. The next section of this chapter explains how, from the first interview, talks were examined for codes and memos were maintained. Through a constant-comparison model, open codes conflated into focused codes. Abduction helped to identify areas that required further exploration. When new questions emerged, previous interviewees were contacted and asked for input. This new information was added to prodigious notes. Reflection on the collected data aided theoretical sampling. Gradually, certain focused codes turned into categories, and categories established a framework. Analysis was shared with the respondents to help ensure their voices were heard correctly. The result was a core category, constructed by both the researcher and the participants, which provided evidence of how professional judgement served as a Heuristic Assessment when determining final report card grades.

2.8 The Nature of Qualitative Research

Section 2.3.3 discussed how GTM started the qualitative revolution (Denzin and Lincoln, 1994b). What is qualitative research, and what is it meant to do? It is not that quantitative research is all about numbers and is associated with the positivist realm, while qualitative research uses words and is aligned with interpretivist thought. After all, a study can use both numbers and words and still needs to define its guiding paradigm. In the present constructivist study, we are looking at how the participants interpreted student achievement by evaluating evidence with provincial criteria, and using personal judgement to assign a symbolic percentage grade. Since there is a mix of numbers and interpretation, is the evaluation procedure qualitative or quantitative, or can it be both?

Whether a study is qualitative or quantitative raises assumptions about the epistemology and methodology, but both approaches are better understood by what methods were used and to what end. Generally speaking, qualitative research parses smaller samples than quantitative research, to answer what, how, and why questions concerning a phenomenon (Green and Thurgood, 2004; Creswell, 2012). In order to understand what social processes mean to a group of people, it is necessary to engage those who are involved in an intersubjective dialogue (Schwandt, 1999; Morse, 2006). How to best interpret data is up to the qualitative researcher. Since the methodology guides the methods, what makes a project qualitative is also tied in with a study's epistemology and theoretical stance.

2.8.1 Constructivist Grounded Theory and Qualitative Research

Since it mostly works with words and a relatively smaller sample, GTM is generally regarded as a qualitative approach. Again, this is not to say that numbers cannot be part, but applying methods such as constant comparison and theoretical sampling are qualitative in nature. As mentioned in Section 2.3.3, GTM suggest a (post)positivist leaning (Charmaz, 1990). Charmaz applied the constructivist paradigm to GTM, thus revising the methods, to establish the qualitative hybrid of CGT. Some have expressed concerns that if methods are meant for one paradigm, they are eroded if moved to another (Greckhamer and Koro-Ljungberg, 2006). However, if the methods respond well to another paradigm, there is no reason not to proceed (Coffey and Atkinson, 1996; Annells, 2006). Charmaz advised,

> "Treat the research process *itself* as a social construction; scrutinize research decisions and directions; improvise methodological and analytic strategies throughout the research process; collect sufficient data to discern and document how research participants construct their lives and worlds." (Charmaz, 2008b, p.403)

Throughout the research, attention was paid to the social construction of evaluation and how the methods could capture insight. This chapter provides a detailed account of how data were gathered and analyzed, along with how research decisions were made to suit the circumstances. The result is a reconstruction of how informed professional judgement becomes Heuristic Assessment in practice.

2.8.2 Quality Assurance In Qualitative Research

In regards to quality in a constructivist study, if anyone's interpretation of the world should be respected, what prevents an "anything goes" approach to research? As Schwandt explained, "we must learn to live with uncertainty, with the absence of final vindications, without the hope of solutions in the form of epistemological guarantees" (1996, p.59). Still, there is a distinct difference between "anything goes" and "good social inquiry." Good social inquiry can be judged by others; it is the reader who decides whether or not good research has been conducted (Angen, 2000). It is not that a study must convince all readers it is the only correct interpretation, but that the work must be trustworthy. Proper gathering and understanding of data can be recognized by the community of researchers. Ideally, the insights in this study could help others better understand evaluation practices, and/or provide a means to look at another phenomena differently, hence adding to the conversation and to be considered efficacious (Mishler, 1990).

To ensure quality in CGT, Charmaz recommended "credibility, originality, resonance, and usefulness" (2006, p.182-183). These points are more guiding ideals than criteria (Schwandt, 1996). Credibility goes back to trustworthiness (Patton, 1999; 2001; Charmaz and Bryant, 2010). Originality involves taking a fresh look at old issues as well as an attempt to introduce new ideas. Resonance is achieved through a rich data collection and thoughtful analysis. Usefulness keeps in mind how the results could serve the community. All four of these points include a democratic approach to co-creating knowledge and working with the

participants throughout the research process. It is also hoped that this study will foster what Schwandt (1996) called critical intelligence: "It requires not simply the ability to understand strategies and implement procedures but also willingness and ability to debate the values of various ends of a practice" (p.69). This practical wisdom is what makes research worthwhile.

It should also be noted that the meaning of terminology in this study was not taken for granted. To leave out such details is to make assumptions that the reader will know the meaning, and important points could be lost in the interpretation. Being clear is essential to quality research. The researcher also should not be limited by traditional definitions, and expand on what is meant by different terms. Misunderstanding can never be completely eliminated in qualitative research. However, providing insight, not definitive explanation, for social processes is the point of such studies (Schwandt, 1999; 2000).

2.8.2.1 Validation in Qualitative Research

Another major concern for qualitative researchers is how to demonstrate rigour and validity. It has been a long and ongoing debate (Miles and Huberman, 1994; Coffey and Atkinson, 1996; Creswell and Miller, 2000; MacDonald and Schreiber, 2001; Lomborg and Kirkevold, 2003; Neuman, 2006; Erickson, 2011; Denzin and Lincoln, 2011; Miles, Huberman, and Saldaňa, 2014). Furthermore, the postmodern perspective has made the concept of validity problematic for all paradigms (Lincoln, Lynham, and Guba, 2011). As previously stated, since methodology drives methods, there should be an epistemological connection,

particularly evident in how the methods provided the means for a valid conclusion (Greckhamer and Koro-Ljungberg, 2006). Positivist and post-positivist studies, qualitative and quantitative alike, rely on factors such as internal and external validity, reliability, and objectivity to justify findings. In contrast, constructivists, who do not recognize an external world that can be viewed objectively, believe the researcher fuses external and internal worlds via interpretation of co-created, subjective knowledge with the participants. Since the situated experiences of participants tend not to repeat, interpretive patterns replace reliability (Denzin and Lincoln, 1994a; 2005; 2011; Guba and Lincoln, 1994; Rolfe, 2006b; Holstein and Gubrium, 2011). Therefore, instead of using quantitative-oriented ideas like reliability, rigour, and validity, qualitative researchers think more in terms of validation. For instance, constructivists attempt to base conclusions on trustworthiness and authenticity, established through a detailed account of the dominant idea in the participants' voices, with an explanation on interpretation (Guba and Lincoln, 2005). Since this study makes proposals on how to improve the evaluation system in Ontario, the reasoning should be sufficiently convincing and presented in a credible manner.

2.8.3 Qualitative Research Summary

According to Morse (2006), "Qualitative researchers sit on the fringes of research, but remember that it is on the fringes where the greatest advances are often made" (p.403). However, being on the fringe also means that conventions are not always clear, and a researcher must balance creativity and recognized procedures for his or her work to be accepted by the community. Corbin (2009)

93

wrote, "Analysis should be relaxed, flexible, and driven by insight gained through interaction with data rather than being structured and based on procedure" (p.41). Qualitative research is loose enough to allow the researcher to explore, while maintaining conventions that can be validated by the community of researchers (Preissle, 2006). The researcher should have room to maneuver, but procedures can help organize data and should not threaten researcher imagination. The next two sections provide a narrative of the procedures that were used in this study, following an adherence to the theoretical framework that has been laid out.

2.9 Constant Reflexivity

The exigencies of research demand a significant amount of reflection. As previously discussed in Section 2.3.7.4, reflexivity allows the researcher to think critically about data and how it is being interpreted, and about the researcher's relationship to the study (Charmaz, 2006; Jootun, McGhee and Marland, 2009; Lincoln, Lynham and Guba, 2011). Consequently, there are research activities that begin at the start of the project, and continue for the duration. Constant reflexivity, combining thinking and action, was needed to complete this study. For instance, continuous consulting of the theoretical and subject literature as well as upholding ethical standards was required. Also ongoing throughout the project was theoretical sampling, memoing, mapping, comparison, and abduction. It is not that one action must follow the other, or that they repeated in a particular pattern. They are ubiquitous principles that were part of all the stages of the research.

2.9.1 Research Ethics

Research ethics address a range of important issues, from site access to treating participants as partners in the research process. Fortunately, this project was not confronted with any access issues. For instance, it was not necessary to gain permission from institutions to speak to the people within. Furthermore, no respondent was considered to be a vulnerable person, i.e., someone who "is in a position of dependency on others; or is otherwise at a greater risk than the general population of being harmed by a person in a position of trust or authority towards them" (*Criminal*, 1985, 6(3)1a-b). Nevertheless, a researcher must always keep ethical principles in mind. If one is careless with something as important as research ethics, then questions may arise over that person's ability to properly apply methodology and methods. As Shaw (2003) stated, "Naivety about ethics is itself unethical" (p.11). An intrinsic approach to ethics, i.e., applying common sense, was sufficient for this type of study (Simons and Usher, 2000; Sikes, 2004; Cannella and Lincoln, 2011; Lincoln, Lynham, and Guba, 2011). The most important action when working with respondents is to simply be honest (Miles, Huberman, and Saldaňa, 2014). These simple but vital guidelines were followed throughout the study.

2.9.2 Informed Consent

As per the guidelines of *University of Nottingham Code of Research Conduct and Research Ethics* (2010), this study followed the basic principle of informed consent. Once the university had granted approval for the gathering of data, respondents were invited to contribute to the study by consenting to an

interview. It was explained that this information would contribute towards a PhD

thesis in Education. Semi-structured interview questions were provided at least

twenty-four hours prior to the talk so participants had an opportunity to prepare.

Permission was granted, orally and on a Participant Consent Form (see Appendix

Two, p.374), to audio record the interview. It was also explained that follow-up

questions would be asked as the study progressed. These questions also gave the

opportunity to update participants, where appropriate, on the progress of the study

and how their informed consent was contributing to the developing construction

(Shaw, 2003). At the same time, the participants knew that they could withdraw

from the study at any time.

2.9.2.1 Power Relationships with Participants

Informing the participants that they could leave the study without prejudice

was one of the many efforts to ensure a balance of power. Another criticism of

Glaserian GTM was the relationship between the researcher and the participants

was not taken into account, justified on the grounds of positivist objectivity.

However, to treat participants as external embodiments of data is unethical. A

researcher must be aware of such power relationships (Smyth and Williamson,

2004; Charmaz, 2005; Hood, 2007). Mills, Bonner and Francis (2006a) advised

researchers to think of respondents as helping with "data generation as opposed to

data collection" (p.10). To show respect, it is important to establish rapport with

the participants (Charmaz, 1995a; 2006; Toma, 2000; Fontana and Frey, 2005).

This does not mean that the interviewer and interviewee must be friends, but the

exchange should be friendly. After all, both parties must trust one another to

create accurate data. The interviewees need to be at ease and open to sharing ideas (Kvale, 1996). Providing questions prior to the interview was one way to help establish a comfortable interview environment.

2.9.2.2 Anonymity, Confidentiality, and Privacy

To protect participants from any potential physical, emotional, or professional harm, their identities were disguised and their schools not named. Professional harm is perhaps the only hypothetical threat in this study. Responses sometimes contained negative opinions on assessment and evaluation practices in the province. If an educator was to link a disagreeable opinion with a specific participant, it could damage that respondent's career. Furthermore, stakeholders, such as guardians/parents, may be bothered by the gaps between policy and practice discussed in the study, including the nature of shared and shadowed practices. Anonymity, confidentiality, and privacy are all part of greater ethical issues of a study (Miles, Huberman, and Saldaňa, 2014).

Respondents were promised confidentiality, indicating that only myself, and possibly the study's supervisors, would have access to the data. For anonymity, they were given the opportunity to choose a pseudonym, to have one picked for them, or to use their real names. Since the findings were grounded in the voices of the respondents, sometimes a participant will want a real name used (Grinyer, 2002). Two interviewees chose this option. However, in case anyone later changed his or her mind about the choice, the naming choices of each participant have not been identified (Forsey, 2012). In addition to the two

participants who used their real names, eleven chose a name and eleven asked that one be chosen. No pseudonym matched the real name of another participant.

2.9.2.3 Data Storage

To ensure the privacy of the participants' identities and opinions, all data were protected. As there was extensive use of electronic means to gather and store data, extra care was needed as this type of research is relatively new and is establishing new ethical principles (Meho, 2006). For example, each participant was provided with an individual electronic folder using the file-sharing program *Dropbox*. This service allows copies of interviews in MP3 format, transcripts, and other documents, to be privately shared. The ability to easily share data with people all over the world was a major asset. It simplified the logistical issues of getting information back to the participant for confirmation. However, this service does pose a privacy risk should a third party infiltrate one's folder. Therefore, care was taken avoid including any personal details in the digital files. The folder, and files within, used the pseudonym or first name of the participant. Names were not used in the interviews, so this detail was not recorded or transcribed. Finally, participant consent forms were only kept in a single hard copy in a secured location. Overseas participants had to electronically send the completed form, and then the digital copy was deleted after printing.

All soft copy files were kept on password-secured laptops. Due to the concern of losing data should the computer be lost, stolen, or corrupted by a virus, the *Dropbox* account synced the files on a home and work computer. In addition,

periodic backups were placed on an encrypted pen drive and kept in a locked cabinet. In addition to the consent form, a research journal and diagram drafts were maintained in hard copy and safeguarded.

2.9.3 Subject and Theoretical Literature Review

As part of the study's proposal, there was a literature review on teacher (professional) judgement in Ontario. Little specific research was found. This realization was looked upon favourably, as it meant it was a topic that needed investigation and increased the likelihood of making a contribution to the field of knowledge. GTM has long been useful for new areas of research (Stern, 1980). However, the proposal did not commit to a specific methodology as the research questions needed to be better defined in order to choose the most appropriate approach (Silverman, 2005). Therefore, the beginning of the study focused on the theoretical literature to get a better grasp on epistemology and methodology, as well as issues concerning proper methods. This investigation led to learning more about CGT, and recognizing it as the ideal way to guide the research. It was also noted that this methodology allowed for gathering and analyzing data before an exhaustive review of the subject literature. As previously discussed, the grounded theorist does not seek to prove what is already written, but to use abduction to form original conclusions (Suddaby, 2006; McGhee, Marland and Atkinson, 2007; Payne, 2007). *Growing Success* and its source material was used, in conjunction with the theoretical literature, to help to refine the area of investigation and the interview questions. After the analysis was completed, reading subject literature replaced theoretical literature. To summarize, the literature review is included

under constant reflexivity because there was always one source or another being consulted to help navigate the research.

2.9.4 Theoretical Sampling

Simply put, theoretical sampling is the decision process of a grounded theorist to determine what to do next in a research project. It is arguably the most important aspect of GTM, as all other research steps can be traced back to some form of theoretical sampling. Theoretical sampling also separates GTM from other types of qualitative research (Opie, 2004b; Charmaz, 2012). Thompson (1999) defined it as "tentative theoretical jumping off points from which to begin theory development" (p.816). Charmaz (1990; 2006; 2014) stated that it starts after a researcher has a better sense of the direction of the research, and Clarke (2003; 2005; 2007; 2009; Clarke and Friese, 2007) indicated that it begins with initial coding. I believe theoretical sampling encompasses the very first step of the project, and continues until its completion. Theoretical sampling should include pragmatic thinking about purposeful sampling, whether or not to consult certain literature, and constructing the nascent research inquiries. It also refines research, such as identifying the qualities a respondent should possess, changes interview questions to explore new territory, and can direct a researcher to ask new questions to previously interviewed participants (Cutcliffe, 2000). As sample size is relatively small in a GTM project, theoretical sampling helps to ensure accuracy and it builds and organizes what is gathered, and encourages theoretical sensitivity, until sufficiency has been achieved (Miles and Huberman, 1994; Pidgeon and Henwood, 2000; Schwandt, 2000; Draucker, et al., 2004; Suddaby,

2006; Bowen, 2008; Willig, 2008; Charmaz, 2012; Hoare, Mills and Francis, 2012).

Documenting theoretical sampling can be challenging (Charmaz, 2000; Wuest, 2001). After all, theoretical sampling incorporates all data interaction (Milliken and Schreiber, 2012). Throughout the rest of this chapter, specific examples of applying theoretical sampling are provided. Offering them in a more chronological order should make more sense to the reader. However, it must be noted that theoretical sampling, in some form or fashion, is behind all steps taken.

2.9.4.1 Follow-Up Questions

Theoretical sampling also guided follow-up questions. Because follow-up questions started after the first few interviews, and continued throughout the rest of the study, the rationale for this procedure is best placed here. The nature of GTM allows the emergent research process itself to guide data collection (Charmaz, 2008b, 2008c). As the researcher is collecting and analyzing simultaneously, research questions will be revised. Moreover, surprising data may lead to the creation of new questions not originally envisioned. Therefore, a GTM project is strengthened when the researcher can go back to interviewees and ask new questions, including feedback on emerging concepts. Some leads may deserve pursuing, while others do not prove to be fruitful. Still, it is this iterative process that gives grounded theory research its momentum (Hesse-Biber, 2007). As Morse, et al. explained, "Returning to interview key participants for a second or third time is oriented toward eliciting data to expand the depth or address gaps in the emerging analysis" (2002, p.16).

However, it must be emphasized that follow-up questions concerned gathering new data, and were not necessarily about clarifying meaning. A participant cannot be expected to recall what was meant by a statement made weeks or months before. Such evidence from the past, no longer directly accessible, further emphasizes the absence of an objective reality (Greene, 1994). It is up to the grounded theorist to clarify meaning at the time of data collection (Schwandt, 1999). Participants present different selves, just as the researcher is a collection of selves (Rolfe, 2006a). The information provided could be biased and/or have discrepancies, or change if obtained at different times (Sandelowski, 1993; Morse, 1994; Opie, 2004a; Wasserman, Clair and Wilson, 2009). Once the data has been collected, it enters the analysis, so accuracy needs to be checked at the time collection. Therefore, confidence in understanding needs to accompany data gathering to promote better overall interpretation (Caelli, Ray and Mill, 2003).

2.9.4.2 Electronic Data Gathering

Again, since follow-up questions were used throughout the project, it makes sense to discuss them as part of constant reflexivity. Sometimes follow-up questions were asked in person or over the phone. In such cases, field notes were taken. This aspect of procedures will be discussed under Memoing. Most of the follow-up questions were done electronically over email or using services provided by social networking, such as *Facebook*. We will now take a moment to discuss the advantages and disadvantages of electronic data gathering.

Using electronic data gathering eliminated many of the issues associated with traditional research procedures, such as the cost of long distance calls, time of travelling to the participants, etc. (Miles and Huberman, 1994; Kvale, 1996; Strauss and Corbin, 1998; Taylor and Bogdan, 1998; Gubrium and Holstein, 2001; Holstein and Gubrium, 2003; Denzin and Lincoln, 2005; Meho, 2005; James and Busher, 2009; Kvale and Brinkmann, 2009). Conducting data gathering literally across the world necessitated using online methods. When theoretical sampling led to a follow-up question, it could quickly be sent out. Participants had the advantage of replying at their leisure, thus further balancing power relationships. Moreover, data was provided already transcribed for coding and analysis.

Still, electronic data is a different medium from traditional procedures, and must be treated as such (Chase, 2005; Markham, 2005). For instance, there is the absence of traditional social queues. However, participants used more than words to express their feelings. For example, a few used emoticons (e.g., the happy face icon, or using a punctuation combination to illustrate an emotion), and/or wrote in all capital letters for stress. Respondents also used acronyms for abbreviations (e.g., IMO for "In my opinion," or IDK for "I don't know"). In the online world, these are all points the researcher needs to notice (Meho, 2005). Fortunately, the participants in this study were skilled with the written as well as spoken word. As a result, the online responses were as rich and detailed as more traditional means of gathering data. Furthermore, since electronic data gathering was used only for follow-up questions after face-to-face interview in person or over *Skype*, I had a general idea of participant parlance to complement the online response.

Nevertheless, a respondent may try to rush a written response. Fastidious analysis of the field notes and transcripts, especially through constant comparison, to be discussed in a moment, confirmed data were useable. Again, it was vital to promptly check meaning with respondents.

2.9.4.2.1 Asynchronous and Synchronous Data

Online data gathering was relatively straightforward. The challenge, however, was keeping the information organized. Electronic data gathering generates two types of data: asynchronous and synchronous. For instance, when asking a single question electronically, generating a single, synchronous response, the data can be copied, pasted, and dated into another document for coding and analysis. However, on a messenger service such as *Facebook*, both parties could be commenting on and responding to different questions out of order. Consequently, asynchronous data is generated. Extra care was needed to ensure data was organized and clarified in a timely manner when in doubt (James and Busher, 2009).

2.9.5. Memoing

In GTM, memoing is also a fundamental procedure conducted throughout the project. Writing memos is like "having a conversation" with the data (Lambert, 2007, p.255). Just as a conversation is a means to get to know a person better, memos allow the researcher to better understand data. As Stern observed, "If data are the building blocks of the developing theory, memos are the mortar" (2007, p.119). A memo can be about anything relating to the research (Pidgeon

and Henwood, 2010). Furthermore, memos can directly contribute to the first

draft of a study (Glaser, 1978; Miles and Huberman, 1994; Charmaz, 2004;

Clarke, 2005; Lofland, Snow and Lofland, 2006; Birks, Chapman and Francis,

2008; Corbin and Strauss, 2008; Birks and Mills, 2011).

2.9.5.1 The Research Journal

From the beginning of the study, a research journal was maintained.

Because of the multiplicity of its use, it deserves its own description. At the start,

it was used to map out initial ideas, courses of actions, and helped serve as an audit

trail (Cutcliffe and McKenna, 2004; Annells, 2006; Milliken and Schreiber, 2012).

It also recorded initial reflections after the reading of theoretical and subject

material. It was through these reflections that promoted theoretical sampling, such

as decisions about everything associated with methodology and methods, such as

the choice of CGT, to purposeful sampling, and pondering situational mapping

(see Section 2.11.1.2). Hence, the research journal in itself was a form a memoing

(Rennie, 1998; 2000; McGhee, Marland and Atkinson, 2007; Birks, Chapman and

Francis, 2008). The journal provided a historical account of the study and helped

strengthen the writing of the thesis (Birks, Chapman and Francis, 2008; Delamont

and Atkinson, 2010; Miles, Huberman and Saldaňa, 2014; Mueller and

Oppenheimer, 2014).

2.9.5.2 Memoing Procedures

Memos do not need to take a particular form, although there are suggested

styles for the purposes of organization. For example, Corbin recommended dating

each memo (Corbin and Strauss, 2008). Charmaz (2006; 2012) suggested

memoing pragmatically whatever helps reflect on data, and becoming more

abstract over time. Clarke pointed out that "Inadequate memoing is the major

problem of almost all research projects; scribbled notes are always better than

nothing, and thoughtful memos on the computer are intellectual capital in the

scholarly bank" (2003, p.561). Some researchers have labels for different types of

memos (e.g., analytical, coding, operational, and theoretical), but this study simply

recorded memos along side data when coding (Glaser, 1992; Lofland, Snow and

Lofland, 2006; Birks, Chapman and Francis, 2008). Questions to be asked were

noted; ponderings about codes and drafts were all part of memos. Throughout the

rest of the chapter, examples of memos are provided (e.g., Section 2.10.4).

2.9.5.3 Traditional Field Notes

As previously stated, not all follow-up questions were completed

electronically. If a participant was asked a straightforward question in person or

over *Skype*, the question, paraphrased response, and observations, were recorded in

a memo that could be called a traditional field note (Montgomery and Bailey,

2007). This information could be recorded in the research journal, or kept in a

separate folder. Important details and potential codes were typed up as part of the

regular memoing procedures.

2.9.6 Situational Mapping

Qualitative research projects are thick, especially in the constructivist

approach to situated description. It is helpful to help the reader view summarized

information through charts and tables (Pidgeon and Henwood, 2010; Buckley and Waring, 2013). Moreover, throughout the research project, creating visual illustrations of relationships in data helps the researcher better understand his or her own project and assists with abstraction (Strauss and Corbin, 1990; Miles and Huberman, 1994; Rennie, 2000; Davison, 2006; Miles, Huberman, and Saldaňa, 2014). Although diagramming is not a mandatory step in GTM, it is common and useful to help ground data (Wilson and Hutchinson, 1996; Strauss and Corbin, 1998). The researcher should keep all drafts, and refer to them to make further improvements (Clarke, 2005). With consistent reflection on drawings, the researcher can achieve deeper insight into the data and make ideas more intelligible (Pederson, 2008). Furthermore, when participants are consulted for member checking, it is easier for them to see results in the form of tables and charts than to read through rich descriptions in order to validate the findings (Buckley and Waring, 2013).

As mentioned in Section 2.4, Clarke's instruction on situational mapping was instrumental in this study (2003; 2005; 2007; 2009; Clarke and Friese, 2007). Her advice on situational, social worlds/arenas, and positional maps guided diagramming. She explained that these maps that examine the relationships between humans and non-human structures could be as messy as the social worlds they represent. Therefore, they can be for personal use only, and do not necessarily need to be cleaned up for use in a final product. Nevertheless, the situational mapping conducted throughout this study led to the creation of the numerous charts that appear, illustrating Ontario's assessment policy, how categories were organized, and summarizing major ideas. Furthermore, simple

tables have been provided to keep information organized. Generally speaking, if an author can write down an idea, s/he should also be able to put it in a diagram.

2.9.7 Constant Comparison

Throughout this chapter, constant comparison has been mentioned in passing as a key part of GTM. A datum is collected, analyzed, and compared to other data. The other aspects of constant reflexivity guide this process. As other researchers have put it, GTM is about making "observations about observations" (Wasserman, Clair and Wilson, 2009, p.362). It involves physical work with codes. The researcher, using a computer screen or words written on cards, literally moves words and phrases from various sources and places them side-by-side in the search for patterns. It is labour intensive. Gradually, iterative constant comparison pieces together germane data to identify codes to build categories, which will establish a core category that answers the research questions. We will see examples of constant comparison throughout the rest of this chapter (e.g., 2.10.5).

2.9.7.1 Abduction

Reconnecting with the original spirit of GTM (see Section 2.3.2.1), as well as incorporating work by Charmaz and Clarke (see Sections 2.3.7-2.4), this study ensured abductive reasoning was firmly active in constant reflexivity. Abduction allows the researcher to work with the ambiguity of data (Charmaz, 2010). As previously discussed, abductive reasoning takes all that is known about a phenomenon, and provides the most reasonable explanation. As a result, it

provides confidence in the findings, and evidence that meaning was not forced.

Rennie (1998; 2000) stated abduction works with induction so the researcher can

be creative – but still grounded.

Like theoretical sampling (Section 2.9.4), it is challenging to describe

abduction apart from other research steps. It is present when forming follow-up

questions, writing memos, creating a situational map, and comparing categories.

According to Pierce, "The truth is that the whole fabric of our knowledge is one

matted felt of pure hypothesis confirmed and refined by induction. Not the

smallest advance can be made in knowledge beyond the stage of vacant staring,

without making an abduction at every step" (cited in Sebeok and Sebeok, 1981). It

can help show how the conclusions were validated by asking the right questions

about the data. Abduction encourages the researcher to question what is

happening in the data and inspires a more thorough examination (Charmaz, 2006).

2.9.7.1.1 Member Collaboration

The goal of GTM is not to prove a hypothesis, but to create one (Charmaz,

2004). Abduction also results in the creation of a hypothesis, with the added

feature of using inductive and deductive steps to build confidence in the findings.

Therefore, if the hypothesis is plausible, then the community who helped create it

should also be able to approve it. Dey (2007) warned that we try to confirm our

assumptions when looking at data. As a result, it is better to validate ideas during

an analysis than to discover interpretations problems afterwards (Morse, et al.,

2002). How to validate, the terms used, differ between paradigms. For instance,

Creswell and Miller (2000) associated member checking with positivist/post-

positivist paradigms, collaboration with critical theory, and prolonged engagement

in the field as constructivist. However, the basic description given to all three are similar: the researcher adds credibility and validity to findings by getting to know participants over time while sharing and confirming findings. As a result, this study refers to participant confirmation of findings as member collaboration. It is in the spirit of the balance of power between researcher and respondent. Blending ideas in this manner is a *bricolage* approach to study, "where borrowing seems useful, richness-enhancing, or theoretically heuristic" (Lincoln, Lyhnam, and Guba, 2011, p.100). Using a *bricolage* is also in the spirit of being a postmodernist as it makes the researcher tie together multiple truths into a coherent message (Gubrium and Holstein, 1997). However, the researcher also needs to be careful not to erode methods and ensure there is epistemological justification through reflexivity and explanation to the reader (Greckhamer and Koro-Ljungberg, 2006).

Member collaboration was accomplished in a similar manner as follow-up questions (i.e., in person/*Skype* or electronically). However, this information has been presented separately as it about validating findings as opposed to gathering new information. Since the intention of this type of data collection was different, even if the style of collection is the same, the researcher needs to keep aims and objectives in mind. For instance, there is also the concern that, in some studies, participants are not qualified to judge results. Their voices may have created the data, but if they are unaware of how qualitative researchers analyze data captured in a moment in time with theoretical sensitivity, should their objections count? In the present study, since the generated insight was meant to assist Ontario educators, based on the collegial voices of knowledgeable respondents, it was

acceptable that the participants were consulted for shared reflection on the evolving conclusions (Toma, 2000; Chiovitti and Piran, 2003; Cutcliffe, 2005; Wasserman, Clair and Wilson, 2009; Pidgeon and Henwood, 2010). After all, "grounded theorists represent, but do not attempt to reproduce, the views of participants, and construct a conceptualization of the data that transcends participants' stories" (Milliken and Schreiber, 2012, p.688). It is not that every participant needed to completely agree with every point, but acknowledgement of how the hypothesis could apply to the local situation with the desire that insight could be provided on a particular social world, such as the Ontario Secondary School.

In short, the participants approved the findings, establishing professional judgement in practice as a heuristic device. However, there are other caveats to member collaboration. For example, participants might confirm the interpretation because it is the path of least resistance. Alternatively, they might try to shape the emerging hypothesis to suit a personal agenda (Sandelowski, 1993). Therefore, such collaboration might actually hurt a study's credibility (Rennie, 2000; Delamont and Atkinson, 2010). Again, the nature of the present study accounted for these dangers. As an insider, I was familiar with the assessment and evaluation framework of the province. Trust in the participants was crucial, but background knowledge allowed me to recognize the reliability and validity of the participant's statements (Kvale, 1994). The participants themselves were aware that that school rules vary from school to school. They could be surprised, even shocked, by how different the rules could be, but did not refuse to acknowledge the possibility. Furthermore, GTM's constant comparison procedures, and Clarke's range of

111

variance, both to be discussed in a moment, permitted multiple views to coexist within the same category.

2.9.7.1.2 Accounting for Bias and Establishing Trustworthiness

Constructivists must also cope with additional accusations of bias since they openly claim to be part of the findings, which goes against the grain of research tradition. For instance, Max Weber proclaimed the importance of neutrality in research findings (Christians, 2011). Even Pierce knew that complete confidence in conclusions based on abduction was unattainable (Reichertz, 2007). Likewise, Silverman (2005) advised to anticipate counterarguments, clearly demonstrate constant comparison, and ensure the findings are based on data. Kvale (1994) believed that bias cannot be avoided, but it could be accounted for.

Attempts to ensure trustworthiness in a constructivist study cannot fully account for doubt. As Morse, et al. (2002) wrote, "While standards are useful for *evaluating* relevance and utility, they do not in themselves *ensure* that the research will be relevant and useful" (p. 17, emphasis in original). Despite the presence of an audit trail, consultation with respondents, and constant reflexivity applied to the data, there is always the "uncertainty" as expressed by Schwandt (1996, p.59). Again, all this study can offer is an honest appraisal of the situation. Transparent efforts have been made to show how validation procedures were built into a dependable and trustworthy hypothesis. Furthermore, an explanation of how this hypothesis could be tested, and improve the situation, gives the study transferability to the real world (Sandelowski, 1993; Delamont and Atkinson, 2010).

Transparency was also offered in the form of connecting the study to the extant literature, and the researcher has not hidden behind the data while delineating the study (Chiovitti and Piran, 2003). According to Rennie (1998), "it is coherently possible for a grounded theory to be persuasive in its own right; it may not require the successful testing of hypothesis derived from it in order to convince the researcher's audience of its truth value" (p.113). Alternatively, there is the position of postmodern ironists:

> "they accept that they can never fully justify their judgements to others or to themselves, but nevertheless maintain the belief that they are in the best that are available at the present time… the researcher simply *believes* her project to be the best, at the same time knowing that there is no epistemological substance to that belief." (Rolfe, 2006a, p.9, emphasis in original)

Rolfe elaborated, "ironists do not argue that all judgements are equally *valid*, but rather that all have an *equal right to be heard*" (2006a, p.11, emphasis in original). It is believed that when this study is heard, those familiar with the situation will receive new insight and recognize trustworthiness, thus validating the information.

2.9.8 Constant Reflexivity Summary

From the beginning of the study until its end, I engaged in constant reflexivity. Considering the literature, ethics, theoretical sampling, memoing, comparison, and abduction were critical parts of the project. Together, they represent the thinking behind the actions taken. Examples of how these different parts manifested in various stages of the project can now be illustrated. In the

following parts, we can take a chronological order approach to how the study was accomplished.

2.10 Initial Research Stage

With permission granted from Nottingham, I began gathering data. Purposeful sampling was used to invite participants to contribute to the study. Constant reflexivity helped to vet respondents to determine who was in a position to provide usable data. At the same time, semi-structured interviews became more refined as what ideas to pursue became more apparent. Analysis in the form of memoing and open coding accompanied data collection. This initial stage clarified the direction of the study.

2.10.1 Purposeful Sampling

Constant reflexivity helped to establish nascent parameters for the study. As mentioned in Section 1.4, purposeful sampling helped to identify participants who would be the most helpful to the study. Only teachers of "qualitative" courses were interviewed. Since I am more familiar with these courses and how they are assessed, it would maximize my ability to act as an insider. In addition, since professional judgement has a significant subjective element, focusing on courses based on subjective material also made sense. Still, there is no way to tell at the beginning of a qualitative project how many participants will be needed to achieve the goal of sufficiency (Morse, 1995; Kvale, 1996; Kvale and Brinkmann, 2009; Folsey, 2012). Therefore, the grounded theorist starts data gathering, and uses the different parts of constant reflexivity to make further decisions.

2.10.1.1 Recruitment Procedures

The study enjoyed a relatively easy recruitment process. Gaining access to sites was not an issue as talks did not need to take place at a specific setting or time. To recruit participants, forty Ontario educators were contacted in person, by telephone, and/or online through email or *Facebook*. Basic details of the study, such as the focus of the interview, were provided.

No monetary incentive was offered for participation in the study. This decision was not just financial, but philosophical. There is a debate over how financial incentives impact the information provided by an interviewee (Forsey, 2012). Besides, there were no challenges in obtaining volunteers. The vast majority contacted expressed interest in participating, with only five potential respondents turning down the request due to prior commitments. It proved too difficult to find a mutually convenient time to talk to eight of the potential respondents, leaving twenty-seven people to be scheduled for an interview.

2.10.1.2 Achieved Sample

Theoretical sampling was used to reduce the total number of participants to active Secondary School teachers. Two of the interviewees had recently left the teaching profession. Consequently, they were not too familiar with recent developments in assessment, and were excused from further data gathering. Also, one of respondents who had recently switched from Secondary to Elementary School was also not contacted for follow-up questions, as the assessment process is significantly different between the two systems. Twenty-four participants were

contacted with follow-up questions. This low dropout rate added to the validity of findings (Hodgson, 2004). Unfortunately, one the respondents passed away before the study could be completed. Only data that was confirmed with him was included in the conclusions.

2.10.1.3 Participant Profiles

Of the twenty-four active participants, an effort was made to have a gender balance in case evidence surfaced of a link between patterns in professional judgement praxis and gender (Castellini, 1999). It was also ensured there was a mix of years of experience. That said, most of the respondents received their teacher training after the introduction of Ontario's New Curriculum. This limitation was not seen as a drawback since the New Curriculum is the current educational model in Ontario. Although the study focuses on assessment practices in the province of Ontario, the majority of respondent experience was from the southern half of the province. However, once overall experience is accounted for, schools from across the province, as well as several Ontario overseas schools, were represented.

Table 2.1 contains details about the participants: years of teaching experience, main discipline taught, initial interview date, and gender. Participants with a blended teaching background are labeled accordingly. Next to the date of the initial interview, there is a note on whether or not the respondent was asked to assist with follow-up questions and the reason, if applicable, for the exclusion. As previously discussed, the names used are a mixed of pseudonyms and real names,

Table 2.1: Participant Profiles

Name	Gender (M/F)	Years of Teaching Experience	Main Subject Area	Date of Initial Interview	Involved in Follow-Up Questions?
Miguel	M	27	Business	January 13/2012	Yes – but passed away during data gathering
Oscar	M	4	Humanities/ Social Science	January 19/2012	Yes
Lorrie	F	4	Humanities	January 26, 2012	Yes
Nelly	F	4	Humanities	January 27, 2012	No – left profession
Fred	M	7	Social Science	January 27, 2012	No – transferred to Elementary School
Derek	M	4	Social Science	January 28, 2012	Yes
Eileen	F	7	Humanities	February 1, 2012	Yes
Catherine	F	4	Social Science	February 20, 2012	Yes
Sal	M	6	Humanities	February 25, 2012	No – left profession
Lucy	F	4	Humanities	March 14, 2012	Yes
Lisa	F	4	Humanities	March 15, 2012	Yes
Dirk	M	4	Social Science	April 20, 2012	Yes
Greg	M	6	Humanities / Social Science	April 24, 2012	Yes
James	M	8	Humanities	April 26, 2012	Yes
Corey	M	6	Humanities	May 6, 2012	Yes
Larry	M	5	Humanities / Social Science	May 14, 2012	Yes
Harry	M	7	Social Science	May 16, 2012	Yes
Winnie	F	2	Humanities	May 18, 2012	Yes
Jordon	M	8	Humanities / Social Science	May 19, 2012	Yes
George	M	5	Social Science	May 25, 2012	Yes
Jerry	M	6	Business	May 31, 2012	Yes
Helen	F	9	Social Science	June 8, 2012	Yes
Denise	F	1	Humanities	August 10, 2012; December 11, 2012	Yes
Murray	M	5	Humanities/ Social Science	September 9, 2012	Yes
Oliver	M	3	Humanities/ Social Science	September 9, 2012	Yes
Sally	F	8	Humanities	January 29, 2013	Yes
Smitty	F	31	Humanities / Social Science	February 7, 2013	Yes

but the distinction has not been noted to protect the anonymity of the respondents. It should be noted that Denise was interviewed twice. As a first semester teacher, I wanted to explore how her nascent professional judgement developed over the first six months of her educational career.

2.10.2 Interview Procedures

Interviewing is a common component of GTM. Interviews can capture details that quantitative means, such as a survey, cannot. For example, the decision making of how a final grade is determined must be explained in detail to appreciate the process (Forsey, 2012). Moreover, when taking a constructivist approach, the interaction between the interviewer and interviewee is necessary to create knowledge. In studies such as these where the interviewer is also an insider, there are opportunities in the semi-structured approach for the parties to explore ideas together (Warren, 2001; Flick, 2002). As a novice researcher, there were concerns about conducting good, academic interviews. Fortunately, as an educator, I am consistently engaged in social interactions that aid interview skills (Fontana and Frey, 2005). Not only is counseling students a significant part of the vocation, there are regular parent-teacher interviews. These interactions helped with the confidence of speaking to respondents about their views on assessment. Furthermore, as Allard, et al. (2007) pointed out, "By occupation, teachers are storytellers who share narratives in schools. Those narratives describe the success, but often, address the challenges and frustrations of teaching" (p.302). The types of questions asked were typical of the ones discussed in common professional conversations, providing thick description for analysis.

2.10.2.1 Interview Preparation

To help ensure a comfortable experience for the participants, and to respect their position as co-creators of knowledge, we discussed their preference for a time and location for the interview (Mills, Bonner and Francis, 2006a). I was as flexible as possible and worked around their schedule. Locations varied from my home, to the home of the participants, and occasionally on a school site. It was the intention to provide the participant with a relaxed atmosphere while also providing privacy.

All interviews were recorded with the software program *Audacity* on a laptop computer. The computer was always plugged into an electrical supply, so running out of battery was not a concern. After a brief chat with the respondent to establish report and clarify last minute questions, the participant was reminded of the purpose of the interview, and that it would take approximately thirty minutes. Informed consent was verified. Finally, a quick test of the software was conducted to check the sound level and clarity.

2.10.2.1.2 Traditional and Online Interviews

Although the location of the interviews varied, each talk could be classified as either traditional or online. "Traditional" indicates we were in the same room at the same time. Online talks used the program *Skype*. Using *Skype* was a major advantage, as most of the research was conducted while I was living overseas and most of the respondents were in Ontario. The same preparation was used to set up the time of the interview, with the location being wherever people felt free to talk

and had a reliable connection to the Internet. Nevertheless, they are different types of interviews and must be accounted for.

When conducting a *Skype* interview, the call was made from either my home or work office. Both offered privacy and a good Internet connection. The participants were always in their own homes, which added to their comfort level. *Skype* allowed for a talk that was similar to an in-person experience, and *Audacity* picked up the dialogue and created a clear recording. The only incident occurred in a conversation with Jordan. The connection failed half way through the talk. Instead of rescheduling, I proceeded to conduct a phone interview by putting him on a speaker. Doing so allowed him to talk while I took notes, but there was the loss of social queues. Extra follow-up questions were used to ensure that I understood his meaning.

Otherwise, the two interview types had much in common. In both cases, we could see and hear each other. Therefore, it was possible to watch for social queues (Forsey, 2012). The semi-structured interview allowed for the development of tangents to explore. If something seemed important to the respondents, by the way they phrased something or a facial expression, I encouraged them to talk about it more. Not only did exploring tangents lead to shaping future interviews and generate follow-up questions, it allowed for the study to flow in new directions as categories were created. I do not feel that data gathering or analysis would have been significantly different had all the interviews had been completed in one way or the other as the information was recorded and confirmed in the same manner.

Overall, the interviews went well. Respondents were open to talking about their practices with another professional. As most of the participants worked at various schools, sometimes we found ourselves comparing our experiences. The talks took on a conversational tone where we often surprised each other to learn about different assessment practices, and how these variations corresponded with policy. Consequently, there was an element of professional development, which benefited both parties. As Charmaz stated, "Participating in the research can also give respondents a different framework with which to look at their experiences" (1995a, p.58).

2.10.3 The Interview Guide

As previously stated, the participants were provided with the interview guide at least twenty-four hours prior to the talk. Each interview had at least one different question because constant reflexivity shaped the gathering of data. Also, the semi-structured approach allowed for spontaneous questions. Since data analysis accompanied data gathering, time was required to apply the guidelines of GTM and determine the best set of questions to ask the next participant.

Table 2.2 illustrates the questions that were used in the first interview. The centre column contains the question asked. The left column illustrates the corresponding areas of interest (see Chart 1.3 on p.48), while the right column provides the rationale for the question. Originally, the interview started with closed questions, and worked towards what I thought would be the most difficult question: the participant's definition of professional judgement (Flick, 2002).

Table 2.2: Original Interview Questions and Connection to Research

Area of Interest	Interview Questions	Rationale
Origin of Understanding	Where did you go to Teachers College? How many years experience do you have teaching Ontario Secondary School? What subject(s)?	These questions qualified the respondent. It was a prior assumption that experience was a significant part of professional judgement. Where and how teachers learned their practices needed to be investigated.
Using formative and summative assessment	In general, how do you put together a piece of assessment?	Since assessment pieces need to be connected to the curriculum expectations, the idea was to see how teachers include policy in their daily practices.
Missing Work	What is your procedure for late/missing assignments?	As punitive measures are a polarizing issue in Ontario, I wanted to see if there was a connection between an opinion on the matter and overall assessment practices. Furthermore, local guidelines on the issue could reveal information about individual school culture.
Levels and learning categories	When it comes to levels, what is the different between a 4- and a 4+?	This question was asked to check understanding of the difference between limited, some, considerable, and thorough achievement on assignments.
Attendance/Behaviour	How do you assess the Learning Skills on the report card?	Learning skills belong to the formative realm of assessment. However, a summative assessment is required for the report card. Therefore, the question allows for a comparison between the thinking process behind learning skills and percentage grades, and how student behaviour enters the equation.
Understanding Policy Student Behaviour Administrative Guidelines	How do you determine a report card percentage grade (with attention to the role of the final evaluation)?	Addressing the central research question, this inquiry allows the respondent to explain the decision making process behind deciding on report card grades. Not only does this process call on a teacher's professional judgement, but demonstrates local guidelines as well.
Attendance/Behaviour Using Formative and Summative Assessment Local interpretation of policy	Is there a correlation between the learning skills and the report card grade?	It was a prior assumption that student behaviour was an influential part of evaluation, despite what is stated in policy.
The Fundamental Principles Criterion vs. Norm Referencing Calculating vs. Determining Grades Use of Minimum Marks	Are Ontario Secondary School courses designed to maximize "passing?"	Relating to some assumptions about student behaviour, as well as local guidelines, this question checks to see if the system in Ontario is designed to maximize passing.

The Fundamental Principles Criterion vs. Norm Referencing	Do you find yourself comparing students to help you decide on marks?	In order to help ensure a fair and equitable assessment, this question checks to see if teachers looks at class achievement, past and present, to assist with criterion referencing.
Origin of Understanding Local Interpretation of Policy	Can you think of ways of the way you mark has changed over time?	Examines the relationship between professional judgement and overall experience.
Understanding Policy Student Behaviour Administrative Guidelines	Anything you would like to change about assessment in Ontario?	Allows for an opportunity to address and concerns they have about Policy. It is also an opportunity to explore tangents.
Understanding Policy Student Behaviour Administrative Guidelines	Briefly, what is your definition of "professional judgement"?	By comparing the definition to other answers in the interview, and the ideas of others, it is an excellent opportunity to use constant comparison.

Note how one question could incorporate different areas of interest, such as the

discussion on how the participant determined the report card grade.

2.10.3.1 Question Framing Strategy

Interview questions should encourage respondents to do the vast majority

of the talking (Kvale, 1996; Dilley, 2000; Flick, 2002). My questions tended to be

open ended, yet focused on the area of interest. Kvale (1994) advised the idea is

"not to avoid leading questions, but to recognize the primacy of the question and

attempt to make the orienting questions explicit" (p.156). Interpreting the

meaning of the respondents, and bringing the voices together as a construction of

knowledge, was a challenging task. Therefore, keeping the interviewees on track

was not a way to force meaning, but to increase the likelihood to uncovering

useful data. Also, I was careful to only ask one question at a time, as it can be

easy to imply two different questions in the same sentence (Charmaz and

Belgrave, 2012). If I did not understand something a participant said, a clarifying

question was used (Miles, Huberman, and Saldaňa, 2014).

Interview questions were designed with sensitivity in mind. Ideally, the respondents would prosaically explain their thinking process behind grades. Therefore, questions were designed to help reveal these processes in a way that would keep the interview conversational. For example, use of educational jargon, such as professional judgement, was limited. Even though both parties had regular exposure to such terminology, it does not necessarily mean they share the same definition or level of understanding. Questions were also asked with what Folsey (2012, p.371) called a "knowing naivety" to get the interview to fully explain their actions and thoughts. When assessment terms were used, I asked for the respondent's definition before continuing. This way, when conducting analysis, I could be more confident in the meaning assigned by the participant (Blaxter, Hughes and Tight, 2006).

2.10.3.2 Connection to Symbolic Interactionism

The questions listed in Table 2.2 also focused on "individual experiences, thoughts, feelings, and actions" (Charmaz and Belgrave, 2012, p.351). As discussed earlier in the chapter such as Section 2.3.4, symbolic interactionists look for such meaning and action regarding social processes in the words of the participants. Clarke said, "Structure is action and action is structure and everything is perspectival" (2005, p.113). Since body language can also suggest something about a participant's meaning, I avoided taking notes during the talk to watch for social queues (Kvale and Brinkman, 2009). For example, in the first interview, there was a moment when the participant Miguel was elaborating on the difficulty of converting levels into a percentage grade. His expression revealed a

genuine frustration with the task along with the words, "They [the OME] should make a decision. Either we're going with levels or we're going with percentages." This evidence encouraged me to further explore this issue with other participants and I found that a similar expression of confusion was common. Investigating the issue of confusion would lead to analyzing how the participants solved such problems in their actual practices.

2.10.3.3 Modifications to the Interview Guide

The initial interviews addressed the report card, but I found I was collecting too much data on practices concerning individual assignments. It is difficult to understand the report card process without referring to individual assignments as they inform professional judgement and the percentage grade. Still, some unnecessary tangents were eliminated. For instance, discussing where teachers received their training was dropped. It became apparent that it is an educator's experience on the job that has a much greater impact on the development of professional judgement. Also, talking about how assignments were put together became irrelevant. Since all assessment should be based on the curriculum expectations, this line of questioning was abandoned as it ultimately had little to do with the area of interest. For much of the data gathering, information was collected on how respondents evaluated learning skills and work habits as there appeared to be a strong link to professional judgement. However, it was decided that focusing on the percentage grade only would best serve the study by focusing on this element of the final report card.

The order of the interview guide also changed. For example, opening the talk with a definition of professional judgement, as opposed to leaving it until the end, made more sense. This way the rest of the interview could build upon this definition as a means to compare and contrast how the concept manifested itself in the evaluation process. Although most respondents found it the most difficult question to answer, asking it first also provided the opportunity to get it out of the way. Besides, answering the question was not so overwhelming that it threatened rapport, and respondents already knew the question was going to be asked.

It would be unnecessary to include the list of questions from every interview to demonstrate how the questions changed or personalized for each participant. A comparison of Table 2.2 to Table 2.3: Final Interview Questions and Connection to Research illustrates the evolution of interview questions. Since I was already working with focused codes at this point, to be discussed in a moment, the questions were more direct as I had an educated guess on how the participant would react and respond. Also, my own confidence as an interviewer had increased. For example, at first I was hesitant to speak directly about the role of student behaviour, as I was not sure how to word the question without sounding like I was accusing the respondent of bias. Eventually, I found participants were willing to openly talk about frustrations with student behaviour and possible links with how these views impacted assessment. Plain but honest questions make for good data gathering.

Table 2.3: Final Interview Questions and Connection to Research

Area of Interest	Interview Question	Rationale
Understanding Policy Student Behaviour Administrative Guidelines	How do you define professional judgement? How has it changed over time?	This particular participant, Smitty had thirty-one years of experience. The questions examine her views on what is professional judgement, and how it has changed over time.
Origin of Understanding	How many years of experience do you have teaching the Ontario curriculum?	Focuses on actual teaching experience.
Administrative Guidelines	What is the role of the administration in the report card process?	Open-ended question that invites the participant to discuss matters that come to mind. There is an assumption here that the administration plays a role, but it should as a general rule.
Student Behaviour	What is the relationship between Student Behaviour and the final grade?	The other respondents had confirmed there was a relationship to some degree. The question still allows the respondent to say there is no relationship. However, I assumed that this participant would be able to provide insight.
Using formative and summative assessment	*Growing Success* stated that conversations and observations can be used for summative assessment. How does this point work in real life?	Most of the participants had indicated that only student products could be used for summative assessment by rule of the local administration. I wanted to see what this participant thought of the situation.
Calculating vs. Determining Grades	How do you pinpoint a student's percentage grade?	Incorporates many points, ranging from local guidelines to professional judgement itself.
Levels and Learning Categories	What is the difference in achievement between a student who earns a Level 4 and one who earns a 4+?	A question that did not change much during the interviews as it always generated interesting responses. Most of the participants struggled to provide concrete examples of the student who goes great as opposed to mastery.
Missing Work	How do you address late and missing assignments?	This question was used throughout the interviews. Generally speaking, the participants appeared to limit the impact of punitive measures, but much depended on the local administration.
Understanding Policy	What do you like about the current assessment policy? Anything you would change?	These questions opened the floor to the participant to address anything else she would like to talk about. Could also potentially bring up something not mentioned in the previous questions.

2.10.3.4 Length and Pace of Interviews

Participants were told the talks would take approximately thirty minutes. The average length turned out to be thirty-two minutes. The shortest interview was twenty-one minutes (Jordan), and the longest sixty-nine (Corey). I rarely did more than one interview on the same day for numerous reasons. First, in order to do constant comparison, I preferred to have at least a day to listen to the interview and memo. Second, most of the respondents were overseas, which meant a twelve- or thirteen-hour time difference. Since the respondents choose a time based on their availability, this could mean, for myself, conducting the interview in the early hours of the morning or late in the evening. Combined with working full-time meant that more than one interview a day was just not feasible. Before each interview, I glanced over the questions. Since questions could change slightly in between interviews, I familiarized myself with the current guide to maximize eye contact with the respondent and to watch for social queues. I kept the questions in front of me as a precaution, but I did my best to remain focused on the interviewee and listened closely to their responses. Charmaz encouraged researchers to listen carefully to the respondents, and try "to learn the unstated or assumed meanings of their statements, and shaping their emerging research questions to obtain data that illuminates their theoretical categories" (2004, p.503). To Charmaz, it was essential to be aware of the situation as described by the participant; it determines everything in regards to "meaning and processes" (2004, p.522).

2.10.4 Listening to the Interviews

After an interview, I played back the recording and only listened. Next, I listened a second time and took field notes of the conversation. The interview was listened to a third time to convert the notes to memos and identify potential follow-up questions. The intention was to think about possible codes that explained relationships between the data and the participants (Miles and Huberman, 1994; Josselson, 1995; Star, 2007). Also, I wanted to feel comfortable with how the talk was interpreted to limit the chance of misunderstanding what the participants were trying to express (Carpenter, 2008).

2.10.4.1 Memoing of the Interviews

In regards to memoing while listening to the interviews, Clarke advised to keep them "partial and tentative, full of questions to be asked and answered about the nature and range of particular sets of social relations, rather than being answers in and of themselves" (2005, pp.102-103). The goal was to provide an appropriate description for the data and explore different manifestations of concepts (Glaser and Strauss, 1967; Pidgeon and Henwood, 2010).

Table 2.4: Example of an Interview Memo contains an example of a memoed excerpt from the talk with Oscar, the second person to be interviewed. Note the conversational tone of the memo and the asking of questions. Sometimes these questions could be answered elsewhere in the text; other times it meant designing a follow-up inquiry for the participant.

Table 2.4: Example of an Interview Memo

Interview Excerpt: Oscar	Memo
Me: Okay, so, we're now moving along to the end of the semester, and you're doing the report card there, could you, you've got Student A, could you please walk me through how you're determining that student's final mark? **Oscar:** Well, ultimately, because I use the software *Markbook* to put all the marks in, and I do keep a hard copy, what I tend to do is, once all of the assessments are in, and the final exam has been completed and marked, I input all the grades, exactly how it's been set out stipulated by department standards or by class standards of what I said exactly what is going to happen, and then I look at that grade. More often than not, I find they are generally accurate reflections of the student's performance, and the student's, you know, mark in the class. However, sometimes, there is the student who, at the very beginning, struggles, however, over the course of the semester, tends to get better, and if I see that improvement, and I see a general improvement in that student's grades, or, over the course of the semester, then what I would tend to do is to either weight the, generally weight the earlier assignments a little bit less, to be a little bit more emphasis on most recent, most consistent, and that generally involves jumping that student from perhaps a 78 to an 80. Or a 74 to a 75, or a 76, in order to get them to what I would consider a benchmark.	(Jan 20th, 2012). He is making references to the local administration, which appears to be in line with ideas expressed in *Growing Success*. What is this relationship like? There is a strong sense of record keeping (e.g., *Markbook*) and following department standards. Does the administration keep track of these records throughout the semester? He reflects about what he thinks about a student's achievement, then looks at the computer-generated grades. He claims that he is usually accurate. Does this mean he has spent time over the semester getting to know individual students and their achievement? When he talks about the student struggling in the beginning of the semester, that suggests that his estimate has been high, then discovers that earlier evidence should be replaced with more recent achievement. Doing so increases the student's grade. Is he placing his impression of the student above what has been recorded? When he says, "if I see an improvement," does that suggest he is double checking his impression against the data and following the principle of "more recent achievement?" It also sounds like he is adjusting marks after the final evaluation, which suggests that all not provincial guidelines are being followed. What is the gap between provincial and local policy? Also, what is the typical process in "jumping" marks?

2.10.4.2 Generating Follow-Up Questions

After generating memos, participants were contacted if there were any points to clarify. As previously discussed, whether or not a respondent can comment on what was meant on something that was said in the past is debated in qualitative research circles (Sandelowski, 1993). However, I believe if the respondent is being asked for clarification within a few days of the interview about a procedure they normally carry out in their practices, it is reasonable that they should be able to provide an accurate clarification. Still, follow-up clarification was used sparingly. Follow-up questions tended to seek new data.

Occasionally, follow-up questions were also a pragmatic way to cover material that was over looked during the interview. For example, I forgot to ask Oscar for his definition of professional judgement. When asked he replied, "Professional judgment is the ability for teachers to provide an assessment of a student that is inclusive of but entirely based upon the marked assessments during the course." As other data were collected, I was able to compare this statement with other responses to build upon the interpretation that professional judgement is something that attempts to appreciate the whole situation, including factors that go beyond individual student achievement.

2.10.5 Open Coding the Interview Data

Charmaz referred to open coding as "wrestling with data" (2005, p.510). She advised that the researcher, through iterative means, monitors action and processes as asks what is happening in the data (2004; 2006; 2012; 2014)? What is the participant explaining? What assumptions are made by the respondents? Also, "how does the structure of context serve to support, maintain, impede, or change these actions and statements" (2006, p.80)? In other words, one must carefully reflect about data's relationship to the situation while coding. Therefore, it is important to not rush coding decisions (Charmaz, 1983; Creswell, 2008). To code properly, I had to be patient, systematic, and tenacious as I read through all the gathered data (Lofland, Snow and Lofland, 2006).

Open coding is about looking for possibilities. It is an identification process for keywords that help interpret meaning in the data. A single participant

does not need to use the exact same words as another to convey the same significance (Dey, 1999; Pringle, 2007). After all, participants tended to tell personal stories when providing examples. As Czarniawska (2009) observed, "One cannot repair a machine by telling how it was done, but one can always tell a story about the repair" (p.651). It then becomes the job of the researcher to see past the story being told to the meaning that is being conveyed. Upcoming tables (e.g., 2.5, 2.6, and 2.7) demonstrate the task of analyzing data, identifying codes, and reflecting on it all through memos.

At first, every code has the potential to be a category (Rennie, 2000). Since open coding was applied to memos of what was said in the interviews, it tended to follow more of an incident-by-incident approach as opposed to line-by-line. By incidents, we asked, what are the consequences? How do the participants see themselves within a group? How are the actions intersubjective (Strauss and Corbin, 1990; Miles and Huberman, 1994; Charmaz, 2006; 2012; 2014)? Again, reflecting on symbolic interactionism assisted with coding data (Milliken and Schreiber, 2012). For instance, I looked at how student-teacher and administration-teacher interactions influenced the grading process. It was interesting to see how educators develop ways to cope with the stresses placed on them. Coping involves sometimes obeying the rules, and other times finding ways they can be manipulated. As with all actions, there are always consequences. Coding helps to organize this information. Simply put, "Coding produces knowing" (Clarke, 2005, p.187).

2.10.5.1 Potential *In Vivo* Codes

When participants do use the same word with the same meaning, it can become a useful *in vivo* code. An *in vivo* code is an exact word or phrase used by the participant. It is best when action and processes can be represented by an *in vivo* code as it makes the analysis more grounded. "*In vivo* codes," Charmaz stated, "help us to preserve participants' meaning of their views and actions in the coding itself… serve as symbolic markers of participants' speech and meaning" (2006, p.55). Table 2.5: Example of Open Coding demonstrates open coding in the study from the talk with Catherine. The left column contains an excerpt from an interview where Catherine is defining professional judgement. Since the study focused on professional judgement, without making major assumptions regarding what would eventually be important, potential open codes were noted. To illustrate this process, these notes have been inserted in the right column.

Table 2.5: Example of Open Coding

Interview Excerpt (Catherine)	Open Coding
I think professional judgement basically is, altering a given mark. So for me, when I am doing my assessments I kind of look at outside factors which would be like, do they participate in class? Have they worked hard for their mark? Do they go to Student Success? Do they come for extra help? Do they use the comments and feedback that I give them on specific assignments?	*Professional judgement* *Altering marks* Assessment process is personal Holistic approach (Considering) *outside factors* Student products Conversations, and observations Participation *Working Hard* *Going to Student Success* Seeking Help Importance of effort *Using teacher feedback*

The *in vivo* codes are in italics. Again, at this stage all codes had potential. Constant comparison of data allowed the key terms to emerge.

2.10.5.2 Avoiding Forcing Meaning

Especially at the stage of open coding, it was essential not to overemphasize the importance of a code. Again, it is best to delay committing to a code until one can be confident in the decision. Codes must be allowed to change (Charmaz, 2012). It is acceptable to experiment with multiple approaches during constant comparison (Willig, 2008). Codes should be anchored to explicit statements by the participants, which is why identifying useable *in vivo* codes is essential (Charmaz, 2005).

2.10.6 Initial Research Stage Summary

The initial stage of research established the parameters of the study. When data gathering began, the area of inquiry needed further clarification. I had my own experience and *Growing Success* to draw upon, but the exact nature of the research was undetermined. The strength of CGT is one can go into a research area with a general idea, and allow theoretical sampling and constant comparison to establish a direction. Surprising data, such as how the level of confusion among participants regarding evaluation, offered opportunities to explore. Furthermore, CGT refined the proper qualifications for a participant, and assisted with the interview and coding procedures. With a better framework established, I was able to proceed to the developing stage.

2.11 The Developing Stage

With the initial stage establishing a foundation, more confidence analyzing data, and information from follow-up questions, work on the study continued in a

developing stage. This stage organized data into focused codes. Next, I began

working on provisional categories. In order to validate my own interpretation, all

the interviews were revisited via transcription. By going back over all the

interviews with ideas in mind, I was better able to identify the strengths and

weaknesses of the analysis.

2.11.1 Focused Coding

Open coding, incorporated into constant reflexivity, allowed for a

transition to focused coding (Lofland, Snow and Lofland, 2006). Charmaz

explained, "focused coding is less open-ended and more directed than line-by-line

coding. It is also considerably more selective and more conceptual" (2004, p.508).

Focused coding helps to elevate a long list of open codes into potential concepts.

As data accumulated, patterns could be observed. The most appropriate word to

represent situated codes (i.e., different words used by the participants to express

the same idea) was chosen. In time, focused codes could become categories.

No coding software was used in this study. Software can be useful, but it

cannot replace the researcher (Corbin and Strauss, 2008). As an amateur

researcher, I felt more confident in managing the codes myself to ensure that I

understood how they were being organized. Potentially, the software could have

been arranging data in an incorrect way and my lack of experience would not have

recognized it. Working more "hands on" with the material was a better learning

experience.

Table 2.6: Example of Developing Focused Codes demonstrates the difference between open and focused coding. In the left column is an interview excerpt from Winnie's interview and her explanation of professional judgement. The centre column shows the potential labels established during open coding (i.e., as illustrated in Table 2.5), while the right column proposes how to unite open codes into Focused Codes in the left column. Again, italics have been maintained to reflect *in vivo* codes.

Table 2.6: Example of Developing Focused Codes

Interview Excerpt (Winnie)	Open Coding	Focused Coding
I think professional judgement basically is, altering a given mark. So for me, when I am doing my assessments I kind of look at outside factors which would be like, do they participate in class? Have they worked hard for their mark? Do they go to Student Success? Do they come for extra help? Do they use the comments and feedback that I give them on specific assignments?	*Professional judgement* *Altering marks* Assessment process is personal Holistic approach (Considering) *outside factors* Student products Conversations, and observations Participation *Working Hard* *Going to Student Success* Seeking Help Importance of effort *Using teacher feedback*	*Professional Judgement* (*altering marks*; personal approach) Holistic Thinking (*outside factors*; evidence of learning) Considering student behaviour (*working hard*; seeking help; effort)

2.11.1.1 Avoiding Forcing in Focused Coding

Relatively speaking, avoiding forcing in the focused coding was less of a concern than in open coding. Since there needed to be strong overlap multiple open codes before they were amalgamated, accompanied by detailed memos for why they were being joined, the connection was more obvious. For example, in Table 2.6, there are open codes for both "professional judgement" as a label and a series of actions (i.e., the altering of marks). In the focused code, professional judgement joined the name and related actions. More specifically, in the spirit of

symbolic interactionism, professional judgement was not just something that a participant had, but the actions he or she took as an educator.

2.11.1.2 Coding Follow-Up Questions

At this point in the research, past follow-up questions were also reviewed to see if the growing body of focused codes could be applied. In Table 2.7: Coding A Follow-Up Question, we see a written response from Catherine to the question, "In regards to effort, do teachers tend to label students as 'caring' or 'not caring?'" The left column contains the open codes that were recorded as part of reflecting on the response in the centre column. This information was compared to focused codes collected from other pieces of data analysis, and placed in the right column. The bottom row includes an example of the Research Journal (see Section 2.9.5.1). In this particular example, we can also see a good example of symbolic interactionism: the participant is reflecting on student behaviour, interpreting it, and deciding on next steps. A memo on the thinking behind the coding has also been included.

Table 2.7: Coding A Follow-Up Question

Question: In regards to effort, do teachers tend to label students as "caring" or "not caring?" (Catherine)		
Open Codes:	**Response:** I do agree that we sometimes put these labels on students. I try to make an effort not to pass these judgements until midterms. Usually by then, we can see the behavior that sends clues. This behavior tends to include a great deal of truancies or lates. It also usually means that students are withdrawn during class, not taking notes, not doing work, etc. We cannot jump to the conclusion however that these students do not care about our class as many times there is another reason for the student's behavior such as issues at home. I have had students who are withdrawn because they have to work at night and/or take care of siblings. Usually when I see these red flags, I try to have a conversation with the student to rule out these possibilities. If the student is not interested in the class, then that's when we have a conversation regarding where the student is going and the fact that they will not be successful if they continue down this path. I also generally get the parent, VP, student success teacher and sometimes guidance involved as well.	**Focused Codes:**
Labelling students		Perception of students
Passing judgements		
Awareness of time in semester		Noting behaviour
Noting *behaviours*: *truancies, lates, withdrawn* attitude, *not taking* notes, not *doing work*		
Avoiding assumptions		Avoiding Assumptions
Benefit of the doubt (i.e., *issues at home*, non-school obligations, student prioritizing		Problem solving
Addressing *red flags*: talking about student success, involving administration, parents		

Research Journal Memo: (April 9/2012) As suspected, it is difficult to resist labelling students, especially when they are making it difficult to get work out of them. The "black and white" question was meant to generate a response with possible tangents, and I think there is something interesting here. A teacher cannot help but notice negative actions (i.e., "red flags") such as being late, not showing up, not participating, etc. However, there is also a perception of these actions. This teacher is trying not to let the negative actions create a negative perception. Instead, Catherine seems to want there to be a reasonable explanation for the behaviour (e.g., problems at home, other commitments, etc.) rather than just a lack of interest in the course. Addressing the issue with the student, and involving the administration and parents, could be good strategies. She is demonstrating that she tries to be proactive about such issues. Nevertheless, there could be a pattern here: it is not the negative behaviours that are important per se, but the overall impression the student gives. Could this impression get imbedded in a teachers mind, and play a factor when report card grades are determined? If a teacher feels s/he made an honest effort to improve student effort, and does not succeed, how does that in turn impact the student's grades? On the other hand, if it is discovered that there are legitimate outside factors affecting the student, does this play a role when considering mark adjustments?

2.11.2 Provisional Categories

When putting together focused codes, I began to think about how they could fit into categories. These labels were provisional as there were still more follow-up questions and analysis to accomplish. Nevertheless, the provisional categories followed some basic guidelines. For example, esoteric category names were avoided. A transparent term was used to identify the conditions and consequences (Charmaz, 1990). Scanning through Tables 2.6 and 2.7, we can see how some focused codes were similar in regards to how the participants interpreted student behaviour. For the time being, I worked with two provisional categories: Student Behaviour and Perception of Students. Gradually, these two merged into the category Perception of Student Behaviour (see Section 3.4).

2.11.2.1 Avoiding Forcing in the Provisional Categories

At the same time, I was careful not to "subordinate the voices and press them into the service of a single narrative" (Delamont and Atkinson, 2010, p.672). It also cannot be over emphasized that the researcher is always part of the study's findings. Charmaz stated, "we define what we record as data, yet how we define data outlines how we represent them in our works. Such definitional decisions – whether implicit or explicit – reflect moral choices that, in turn, spawn subsequent moral decisions and actions" (2005, p.511). As Castellani reminded us, "one can never escape one's social context" (1999, p.263). Likewise, Gibson warned, "the novice mistakes the order in which categories have been discovered for a basic social psychological process in the data, when in fact it is their own social

psychological process of discovery that they have mistakenly written into the coding" (2007, p.445). Again, constant reflexivity was needed to check research decisions.

Categories were meticulously examined to ensure the findings were based on the voices of the participants. However, if I am the one choosing the voices to share, how can I gain the reader's trust? Dey said, "categories are not simply generated by data, but through judgement in terms of some cognitive frame of reference by which we make sense of the experience" (2007, p.170). Consequently, the researcher needs to apply a strategy on sensitizing

Table 2.8: Example of a Provisional Category (Student Behaviour)

Provisional Category: Student Behaviour	
Focused Codes	**Memo**
Participation	September 12, 2012: Ontario teachers spend a considerable amount of time getting to know students. Teachers come to know who participates, attends class, completes homework, etc. That said, *Growing Success* instructed teachers not to consider Student Behaviour when conducting evaluations. However, when discussing how they use their professional judgement to determine a report card grade, participants appeared to, at least to some degree, reflect on student behaviour. When providing examples of justifying mark adjustments, participants included at least one instance of a student physically doing something that enters into the thinking process. Therefore, this lower-level category collects descriptions of student behaviours that persuade evaluation.
Following rules	
Following instructions	
Completing Homework	
Punctuality	
Seeking (extra) help	

concepts from active data, explain this data from the point of view of the participants, all the while recognizing that the researcher is also part of the explanation. In Table 2.8: Example of a Provisional Category, the focused codes, noted in the left column, concerning actions taken by students based on the experiences of the participants, labelled Student Behaviour, were joined as a

potential concept. The rationale for this provisional category is explained in the memo in the right column.

2.11.3 Transcribing Procedures

Full transcription of the interviews was delayed until provisional categories were established. At this point, eight had been identified: Confusion, Dealing With Administration, Experience, Perception of Students, Punitive Measures, Professional Judgement, Student Behaviour, and Making Sense. I felt that these provisional categories could be further collapsed. With transcripts, I could go back and reviewed the coding decisions and memos that led up to the provisional categories (Charmaz, 2004; Kvale and Brinkmann, 2009). No third party was hired to do the transcribing. I thought it was necessary to do it myself so I could more deeply reflect on the provisional categories and data sufficiency.

Since traditional and online interviews were recorded in the same manner, there was no difference in how they were transcribed. I listened to and typed up the MP3 recording. I could see what Rennie (2000) meant when he talked about how transcribing deepens understanding, as "the understanding of the whole of the text influences the understanding of a part of it, and the understanding of each part in turn influences the understanding of the whole" (p.484). By comparing the transcripts to established codes, patterns became clearer. For example, it became obvious that Student Behaviour and Perception of Students ultimately discussed the same subject matter, as it was the behaviour of the student the participant perceived. Also, I could see that Professional Judgement was inefficient as a

separate category. The codes that were used for this category were better used as shading to highlight actions taken in the other provisional categories. Punitive Measures and Dealing With Administration appeared to be working as their own categories, although the latter was renamed Administrative Guidelines. There was still the matter of what to do with the focused codes provisionally organized as Confusion, Experience, and Making Sense.

Transcribing was not simply of matter of listening and typing. Decisions needed to be made about other factors that could imbed meaning. For example, at first, I attempted to be as accurate as possible in regards to noting hesitations, inflections, and other sounds picked up on the recording. After ten transcripts, it was decided that this extra step was too time consuming and did not appear to provide added insight into interpretation. The subsequent transcripts only contained what was said in the interview (Hammersley, 2012).

I also noted how challenging it was to maintain word-for-word accuracy when typing a transcript. Entering the wrong word or missing a word could change the meaning of a sentence. Listening and re-listening to the talk meant it could take six-eight hours to type a thirty-minute talk. Nevertheless, the process provided a greater understanding of what the participant was trying to say, especially when compared to the follow up questions. Quality is better than quantity, but spending a quantity of time with the data provided a greater quality of understanding (Charmaz, 2004; Oliver, Serovich and Mason, 2005; Morse, 2007; Alvelson, 2011).

2.11.4 Developing Stage Summary

The developing stage was a major move forward for the study. By the end, I had provisional categories, fortified with focused codes. All the interviews were transcribed and I had ideas about a core category. However, some of the provisional categories lacked sufficiency. More data needed to be collected.

2.12 The Enhancing Stage

With the initial interviews over, and provisional categories established, data gathering was nearing completion. An enhancement stage was needed. There were still a series of follow-up questions that needed answers in order to finish building the categories. Once I was confident in working with four categories, theoretical coding analyzed how they compared to one another in order to establish a core category. Also, a range of variance needed to be applied to the categories to incorporate all the data.

2.12.1 Establishing Sufficiency

Miles and Huberman (1994) warned that the coding process could go on forever because data can always be examined in different ways. Therefore, the researcher needs to know when to stop. With an unlimited number of tangents to explore, focus needs to be maintained (Holton, 2007). This study is about the role of professional judgement in determining report card grades. Since professional judgement can manifest itself in every aspect of teaching, I needed to be careful not to stray too far from the research question. At the same time, the grounded

theorist needs to recognize the opportunity in surprises, and allow the data collection to help direct the study.

There are no specific criteria for sufficiency. Traditionally, it has been known as saturation, but Dey (1999) pointed out that sufficiency is a more accurate term. Abduction taught us that new data could always reveal a surprise. However, when the various questions begin to feel like hypothesis testing because the researcher can anticipate how the participant will respond, sufficiency has likely been reached (Bowen, 2008). Morse, et al. said, "Saturating data ensures replication in categories; replication verifies, and ensures comprehension and completeness" (2002, p.12). The researcher has reached a point of understanding on a particular concept thanks to the dialogue with the participant (Schwandt, 1999; Caelli, Ray and Mill, 2003).

2.12.2 Range of Variance

Researchers using GTM, including CGT, have used negative cases to address data that does not appear to fit in any one category (Glaser and Strauss, 1967; Morse, 2007; Bowen, 2008; Birks, et al., 2009). However, Clarke (2005) pointed out that since all the data comes from the situation, it must fit somewhere without forcing. In Section 2.4.2, we looked at her idea of the range of variance and how it takes CGT further along the postmodern turn. Clarke recognizes that identifying negative cases is a positivist legacy where patterns must be neatly organized and what does not appear to work can be dismissed as dross (see Section 2.4.2). Using a range of variance works well with CGT and accomplishes the long-held goal of "find[ing] ways to apprehend and re-present these different

representations to achieve the 'fuller knowledge' that advances knowledge and influences practices" (Sandelowski, 1993, p.3).

According to Gibson, "Critical grounded theory might be able to make general predictions or statements that cover all groups of people, however it would also have to remain sensitive to variable differences within groups" (2007, p.449). Using a range of variance gives the researcher a tool to work with these differences. Differences help to highlight significance in data (Willig, 2008). Clarke explained,

> "specificities of meaning *within* particular situations are important to grasp while also grasping and using theory and other research to enhance our understanding. We need to grasp variation within data categories, range of variation within data, complexities, contradictions, multiplicities, and ambivalence(s) manifest individually, collectively, and discursively." (2005, p.27)

For example, analyzing how respondents defined professional judgement, and provided examples of its use, showed how it manifested itself in different arenas and helped to form the different categories discussed in Chapter Three. In Administrative Guidelines, professional judgement is recognizing that part of the teaching profession was to follow the directives of one's superiors; in Punitive Measures, professional judgment was to use one's experience to determine whether or not negative marks should be applied and why; in Perceptions of Student Behaviour, professional judgement was interpreting the actions of students and reacting accordingly; and in Holistic Achievement, professional judgement was the ability to analyze a student's most consistent level of achievement within

the given the situation and pinpoint a percentage grade on the report card. The range of variance also works within a category. For instance, Administrative Guidelines (see Section 3.2) was a vast category because it reflected how different evaluation practices were from school to school. In some schools, administrations can be quite strict in enforcing evaluation guidelines, including directives that go against *Growing Success*. At another school within the same board, a participant could feel that s/he was given academic freedom to evaluate. It is common that evaluation software was used to determine grades, but whether it was a tool or something the participant must mimic when doing the report card also provided a range of responses. Also, some respondents reported that the local administration would unilaterally change grades on a report card, while others had never heard of such a practice. In short, some participants felt free to use their professional judgement, but others felt the administration was the antithesis of the concept. These examples are stark differences, but they all relate back to how an administration manages the school. The various actions can co-exist within the same category, just as they co-exist in the real world. Table 2.9: Range of Variance in the Administrative Guidelines Category (see Section 3.2) provides further details of the focused codes that were used to build the concept. The left column again shows the ever developing focused codes and the right column is another example of a conceptual memo explaining, to myself, how the codes are linked together in the range of variance.

Table 2.9: Range of Variance in the Administrative Guidelines Category

Focused Codes	Memo
Administrative Adjustments Computer Dominance Confusion Freedom Justifying Grades Meeting Expectations Opposes professional judgement Passing borderline students School Culture Self-taught Trust	March 15, 2013: The conceptual category Administrative Guidelines concerns what the participants think they should do as teachers when it comes to evaluation. These actions are tied to the culture of the school, which is established by the local administration. Some participants reported prevalent administrative oversight, while others expressed the presence of academic freedom. Some even felt that Administrative Guidelines were the antithesis of professional judgement in the way they can unilaterally change grades and have strict enforcements of meeting certain expectations such as relying on computer-generated averages to determine report card grades. Either way, participants generally agreed with the local rules as they felt that, while they might not always agree, the structure is needed in order to have a sense of evaluating properly. For example, it is expected that students should be able to pass the course if the correct actions are taken. Should a student fail, it should not come as a surprise to any of the concerned parties (e.g., teacher, administration, student, parents/guardians). Ideally, there is a mutual trust between participants and the administration to help ensure a good work environment.

The core category (see Section 3.6), Heuristic Assessment, also took advantage of the range of variance. As it will be discussed in detail in the next chapter, Ontario schools are a place of shared and shadowed practices. Applying professional judgement, regardless of the category, requires knowledge and understanding of how to navigate this world in regards to evaluation practices. Although assessment and evaluation is meant to be fair, transparent, and equitable, it is actually challenging to be all three at the same time. Using the principle of acting in the best interest of the student, even if certain actions go against the Fundamental Principles, tended to win out in the end.

2.12.2.1 Validation of the Range of Variance

The researcher cannot use the range of variance as a matter of convenience to put contrary codes into a category and claim that the category is sufficient and conflicting data merely reflect differences in the real world. In the spirit of establishing trustworthiness, the researcher must ensure that the data within in the range works towards a common goal. Again, it needs to be shown that although codes may conflict, they are referring to a common concept. Since the social world is one of ongoing negotiations, demonstrating how opposing views lead to action and reaction is to show the dynamic nature of the real world. If the findings ring true with the community of readers, then validation has been accomplished.

It should also be noted that the ends of the spectrum could potentially be the result of participants inaccurately describing their situation and/or a misreading of their meaning in the analysis. In this study, the likelihood of either happening was greatly reduced by my own insider knowledge and clarifying questions. As a co-constructor, I could reflect on my own experiences and use the *bricolage* approach to analysis (Denzin and Lincoln, 2005; Richardson and St. Pierre, 2005; Kennedy, 2009; Kincheloe, McLaren and Steinberg, 2011). Therefore, I could personally validate the range of variance, and trust my own instincts in constructing the categories. Again, if I have provided insight, informed readers will recognize the presentation as trustworthy.

2.12.3 Confirming Categories

Once a provisional category appeared to have sufficiency, it could be confirmed as an official category. The titles given to the categories, Administrative Guidelines, Punitive Measures, Perceptions of Student Behaviour, and Holistic Achievement, sound similar to the previously-identified arenas of professional judgement (i.e., Understanding Policy, Student Behaviour, and Administrative Guidelines). It must be emphasized that these arenas in *Growing Success* are not pre-conceived categories. The four identified categories are conceptualizations of the data, based on the voices of the participants. The arenas were used as a starting point to generate discussion with the participants, which is acceptable in a CGT (Kennedy and Lingard, 2006). Since the data stemmed from discussion of these general topics, it was not surprising the titles given to the conceptual categories would be reminiscent of the arenas. It is the data analysis within the conceptualized categories that shows they were not pre-conceived and forced with thematic description.

In short, the remaining provisional categories of Confusion, Experience, and Making Sense were distributed between the four categories. Combined together, they also aided in establishing the category of Holistic Achievement. Punitive Measures was a noticeably smaller category, but the information it contained stood alone. A category can be relatively smaller but still distinguish itself from the others (Pidgeon and Henwood, 2010). The data used to build these categories, with a strong emphasis on the voices that helped create them, is discussed at length in the next chapter.

Table 2.10: Perception of Student Behaviour provides a list of focused codes (left column) and a memo (right column) to illustrate the main idea of the conceptualization and how it impacted the final report card. Further details are provided at length in the next chapter. As stated earlier, the category was formed by combining focused codes from the provisional categories Perception of Students and Student Behaviour. With this approach, the conceptual category not only identifies the types of student behaviours the participants noted, but also how the respondents reacted. Should a student display desirable behaviours, the respondent would be more likely to find a way to increase the student's final grade. Students who did work, but did not have a good attitude, would still get the mark they earned, but would unlikely receive any additional marks. A student, according to the participants, would not fail a course simply because of behaviour.

Table 2.10: Perception of Student Behaviour Category

Perception of Student Behaviour	
Focused Codes	**Memo**
Attendance Attitude Effort Following Class Rules Going the Extra Mile Impression Improvement Outside Factors Participation Respect Rewards Working Hard	May 31, 2013: Although *Growing Success* advised to not consider student behaviour as part of achievement, the participants found this directive very difficult to follow in practice. Behaviour, including attendance, participation, and respect, cannot be ignored. Generally speaking, students who are deemed to be hard working or "go the extra mile" will benefit from a more generous evaluation. At the same time, the participants recognize that there can be outside factors, natural shyness, etc., that can interfere with achievement. Therefore, the participants make an effort to get to know their students, and they expect effort in return. By showing respect, they hope to receive respect. Hopefully, all students demonstrate some improvement during the course. It must also be noted that students who have poor attitudes, but finish the work, receive the grade they earned. However, if a teacher senses that they were capable of doing more, they are unlikely to receive extra percentage marks on the final report card. Students who do the best they can, in the eyes of the teacher, are more likely to be rewarded.

As part of Member Collaboration, the basic framework of the conceptual categories was discussed with the participants. Generally speaking, they agreed with the idea behind the categories. Since conceptual categories were abstracted from the voices of all the participants, respondents tended to see how the idea applied to the situation as a whole in regards to the evaluation process. Although some points were not applicable to individual circumstances, they could see how the category was feasible.

2.12.4 Comparing Categories

With the conceptual categories established, they could be compared to one another to establish a core category (Charmaz, 2004; 2006; 2014). Just as one should not rush coding decisions, there is no need to identify the core category early. In fact, the researcher can experiment with multiple core categories until one wins out (Corbin and Strauss, 2008). When looking at all the categories at the same time, just as they consistently work with and against each other in the real world, the cause and effect needed to be determined and what was "conjuncture… There ought to be ways for us to understand how human events and meanings, actions and intentions, are chained over time, as slippery and disorderly as they may be" (Miles, Huberman, and Saldaňa, 2014, p.222). When the four categories are compared, what patterns emerge?

The next chapter will provide a more detailed analysis. In the meantime, it can be said that the four categories are a good example of the negotiations that happen within the evaluation process. In the social world of the school, there are

grey areas of shared and shadowed practices where everyone wants to try to be fair

and equitable, but not always transparent. For the participant, evaluation can be

emotional as they care about preparing students for the real world. Consequently,

they try to maximize student success by justifying the highest grade they can,

including passing borderline students. At the same time, there are the local rules

enforced by the local administration that must be incorporated into a participant's

practices. With so many competing factors, the respondents ultimately needed to

establish a sense of stability by creating their own set of rules based on a

personalized approach to professional judgement. These "rules of thumb" were

termed Heuristic Assessment.

2.12.5 Enhancing Stage Summary

By the time the enhancement stage was completed, all the categories had

been confirmed. Using a range of variance allowed for all the data to be

incorporated, and it was felt sufficiency had been achieved. There was also the

added benefit of reducing the number of categories, which made them easier to

compare. Member collaboration confirmed that the ideas behind the categories

resonated with the respondents. Comparing categories established a provisional

core category that needed to be confirmed. Most of the data gathering and

analysis had been completed. The study was ready to move into its Completing

Stage.

2.13 The Completing Stage

In order to finish the project, the core category still had to be explained to the participants for feedback. Presenting the findings in a similar manner as the categories confirmed that the interpretation provided insight to the role of professional judgement as a heuristic device in determining report card grades. I was then able to turn to the subject literature to determine where the findings fit in the extant field of knowledge. From there, writing decisions needed to be made on how best to present the study. After an editing process, the study was finally written up and was ready for review.

2.13.1 Core Category Confirmation

Strauss (1987, p.36) provided guidelines for a core category. In addition to the previous discussed conditions appearing in the other categories without forcing, "it should be sufficiently abstract so that it can be used to do research in other substantive areas… [and] it should grow in depth and explanatory power as each of the other categories is related to it through statements of relationship." In this study, Heuristic Assessment emerged as the core category. Table 2.11: Summary of the Core Category (Heuristic Assessment) provides a quick overview of this category (left column) and its relationship to the other categories (right column). As previously discussed, after the principles of the core category were determined, they were shared with the participants. Encouraging feedback allowed me to proceed with the chosen label. A more detailed discussion of the care category is contained in the next chapter.

Table 2.11: Summary of the Core Category (Heuristic Assessment)

Heuristic Assessment	
Other Categories and Relationship to Core Category	**Memo: Category Connection to Core Category**
Administrative Guidelines Considering/negotiating local rules for evaluation **Punitive Measures** Considering late/missing assignments as part of the evaluation process **Perceptions of Student Behaviour** Considering observations of students as part of the evaluation process **Holistic Achievement** Considering not only a student's most consistent level of achievement, but considering the student as a person	August 2, 2013: In regards to the role of professional judgement on determining Ontario Secondary School report card grades, the process is best explained as heuristic assessment. As the label suggests, when deciding on a student's mark, a teacher considers whatever they know about the given situation – and comes to the best decision possible. Teachers have personal "check lists" they consult to help with the process of evaluating student achievement. Although *Growing Success* highlights the importance of professional judgement, due to its confusing nature, heuristic assessment is used as common sense solutions that often go against the policy itself.

2.13.2 The Subject Literature Review

The section on Constant Reflexivity stated that there was on ongoing literature review throughout the study (see Section 2.9.3). This review consisted of both theoretical and subject literature, with more of an emphasis on the former. With the analysis complete, full attention could be paid to the subject literature. Delaying the bulk of the subject literature review not only avoided the temptation for hypothesis testing, but also was in itself its own reward. I could see similar patterns in the literature, while also recognizing the originality of my own work. How the present study fit into the literature, and how it is a contribution to the field of knowledge, is discussed in Chapter Four.

2.13.3 Writing Decisions

The pragmatists place much value in a well-written argument (Strauss and Corbin, 1990; 1998; Rennie, 2000; Wolcott, 2009). As previously stated in

Section 2.9.5, memoing is invaluable to GTM researchers because it is not just about generating ideas – but is contributing towards the first written draft of a study. Memos become more and more specific throughout the research process, and raw information is organized into initial report chapters. However, memos need to be reread, transcripts need to be reviewed, and participants need to be re-consulted to transform the words of the participants into a coherent narrative. The aspects of constant reflexivity are still at work as the researcher reviews the report gaps (Charmaz and Mitchell, 1996; Mills, Bonner and Francis, 2006; Charmaz, 2014).

Multiple texts are automatically written in qualitative work. Lincoln, Lynham and Guba (2011) observed, "multiple selves – ourselves and our respondents – of postmodern inquiries may give rise to more dynamic, problematic, open-ended and complex forms of writing and representation (p.124). The study needed to recreate the researched world in written form (Richardson and St. Pierre, 2005). In order to organize ideas, I approached the write up as if I were writing a non-fiction story (Charmaz and Mitchell, 1996; Charmaz, 2006; Birks, et al., 2009). The researcher should use this reality to place the data along a plot line. In the words of Rolfe (2006a, p.9),

> "Writing is the creative process in research… if all our effort is put into Method, nothing is left for writing, and the creative process is stifled… rigid and rigorous application of the scientific method results in 'the inhibition of intuitions' such that '[the researcher] imagination is restrained and even his language will cease to be his own.'"

While writing up the data, I ensured the voices of the participants were dominant. At the same time, as a co-constructor of knowledge, I did not hide behind data and findings. The study is more emic then etic. This presence is evidence in the use of the first person, as well as being honest with the reader about personal viewpoints and interests. Providing such reflection provides background of how the research project came about and demonstrates awareness of prior assumptions. During the analysis, these details needed to be confronted in order to avoid forcing meaning (Charmaz and Mitchell, 1996; Caelli, Ray and Mill, 2003; Mills, Bonner and Francis, 2006a).

2.13.4 Editing and Final Write Up

With the chapters written, editing work began in order to bridge the information together. In order the research to come together as a narrative, the chapters had to complement one another. At the same time, the basic conventions of qualitative and thesis writing needed to be used. The opening chapter was straightforward as it outlined the study as a whole. The present chapter outlined the methodology and methods, while the third provided the analysis. I had to decide where the literature review should be presented. I decided that it would sound better coming after the analysis. Finally, the conclusion focused on what was learned from this study and how it could be applied to evaluation practices in Ontario.

Chart 2.1: Constant Reflexivity and the Research Stages

COMPLETING

Core Category
Confirmation

Writing Decisions

Final Write Up

INITIAL

Purposeful Sampling

Interview Procedures

Open Coding

Ethics

Literature Review

Theoretical Sampling

CONSTANT REFLEXIVITY

Memoing

Situational Mapping

Comparison

Abduction

ENHANCING

Sufficiency

Confirming Categories

Comparing Categories

DEVELOPING

Focused Coding

Provisional Categories

Transcribing Procedures

2.13.5 Completing Stage Summary

Even with the data gathered, there was still a lot of work to do to complete

the study. Confirming the core category with the participants was one matter, but

taking the memos and editing them into a presentable fashion, covering all the

subject literature and writing it up, and bringing the chapters together as a coherent

whole was demanding. In Chart 2.1: Constant Reflexivity and the Stages of

Research, the whole thesis process is displayed as a combination of cyclical and

linear processes, working with one another towards a common goal. The finished

project is here for the reader to review.

2.14 Methodology and Methods Summary

As demonstrated in this chapter, the qualitative researcher is confronted

with numerous hurdles to overcome in order to produce a study that balances

concerns about subjectivity and objectivity, validation and reflexivity, and

epistemology and method. Ultimately, the researcher needs to take a stand on

what feels right. This study chose CGT to unite methodology and method based

on a careful consideration of the research questions regarding the use of

professional judgement in Ontario Secondary School final report card grading.

This approach allowed the voices of the participants to be unified. Data were

gathered via interviews and follow-up questions, and confirmed with the

participants. While the findings only claim to represent one version of truth,

jointly constructed by all those involved, the point of contemporary qualitative

research is not to create a final say on a given matter – but to make original

contributions to the conversation. This contribution is further explained in the

next chapter as the construction of the categories is illustrated to explain the

study's findings.

CHAPTER THREE: THE CONSTRUCTION OF HEURISTIC ASSESSMENT

3.1 Introduction

Like most, if not all, social worlds, school is a place of explicit and implicit rules. Knowing what to do and when to do it ranges from the subtle to the obvious. When it comes to the rules of assessment, *Growing Success* instructed teachers to use informed professional judgement assiduously to combine policy, a clear understanding of achievement and its difference from behaviour, and directions from the local administration. However, since the term professional judgement can be generic, for the purposes of this study, a more precise locution was needed to explicate the process of how the participants take evidence of achievement of the curriculum expectations, and convert it to a percentage grade on the final report card. Educator insight into what happens in school is revealing (Bailey, 2000; Cheung, 2002). Findings suggested that for the participants professional judgement is a heuristic device: a personalized, and simplified, methodology to make report card decisions within the paradigm of provincial standards. This approach was named Heuristic Assessment. Heuristic Assessment, as opposed to informed professional judgement, provides an all-purpose set of rules to help decipher individual achievement and reach a satisfying conclusion on a final percentage grade in a situation with competing, and often confusing, conditions. This chapter elaborates on how CGT was applied in this study to co-construct the findings with the respondents. Identifying the conditions that cause Heuristic Assessment is a contribution to the field of knowledge

because it provides insight into the grading culture of Ontario's Secondary Schools.

Heuristic Assessment exposes the fallible nature of professional judgement, informed as much by symbolic interactionism as it is by student achievement. Expressing achievement as a percentage grade is an overwhelmingly subjective exercise. The teacher is asked to take the concept of achievement, a concept that is fixed to a particular time and place, and compare it to the separate concept of content standards that do not contain cogency. Furthermore, two different schools, even within the same board, could interpret Ontario's wide-ranging policy in ways that are different, but still in line with the policy itself. After all, one's perception of what is fair and equitable for an individual student will differ depending on the situation. Such conditions create a framework of shared and shadowed practices. Shared practices refer to assessment issues that are either common in Ontario schools, and/or are openly discussed at an individual school. Although policy suggested the assessment process must be transparent, there are shadowed directions and decisions administrations and participants prefer not to discuss with students and parents/guardians. Such practices are not significant secrets to insiders; these aberrant practices were known to most participants. To cope with balancing objective and subjective factors, and shared and shadowed practices, the participants demonstrated they developed personal paradigms. They assess in a way that made sense to them, while prosaically explaining to students and parents/guardians, and still be compliant to standards. Although elements of Heuristic Assessment are similar,

there is no universal rule set. Professional judgment is as unique as the participant

who practices it, so there were diverse approaches to reading, and assigning value

to, evidence of learning among the respondents. However, by using a range of

variance, a similarity guiding Heuristic Assessment was noted: the voices of the

participants suggested they assess propitiously. Simply put, the final percentage

grade was generally the highest they could justify when all factors, i.e., shared and

shadowed practices, achievement and non-achievement, were considered.

Furthermore, the participants gave the impression that Heuristic Assessment

turned active, or informed, professional judgement into a last step in the evaluation

process, rather than applying professional judgement as a guiding force throughout

the process. In other words, a common feature of Heuristic Assessment is to take

a look at a grade then deciding if it can be adjusted.

The world of assessment in Ontario directs educators to treat each student

as an individual learner while reporting achievement that corresponds to provincial

standards. Heuristic Assessment allowed participants to mentally negotiate the

realities of assessment and evaluation while applying casuistry in the best interest

of the student. The intention was to take what is known about student

achievement and be consistently fair and equitable, but some decisions could lack

transparency due to the pull of subjectivity. Because of limits to knowledge and

time restraints, evidence of student achievement was usually given the benefit of

the doubt. Heuristic Assessment also helped participants solve antinomy in

guidelines. When these participants were confronted with an assessment decision,

and following one aspect of the local guidelines felt like a violation of another, the participant's resolution was likely to be in the favour of the student.

Denzin and Lincoln (2005) explained that a researcher reconstructs a social world for an audience. To do so is to appreciate the nuances of meaning; what actions means to the actors, how both action and reaction are products of a specific environment, rules governing the environment, and most importantly, how it all comes together as a structure. This study asked, what does the process of professional judgement mean to the actors involved, and what does this concept is reflected in action? The previous chapter explained how CGT established conceptual categories. This chapter examines these categories: Administrative Guidelines, Punitive Measures, Perception of Student Behaviour, and Holistic Achievement. They are conceptualizations of policy in practice, inspired by *in vivo* codes, grounded on the voices of the participants. The respondents narrate the findings by describing and comparing their location situations. Each participant was included in every category. Category subheadings refer to focused *in vivo* codes that helped to construct the category. Charmaz said true grounded theory is used to "conceptualize a problematic process, construct analytic categories from inductive, comparative coding of data, define the properties of the categories, specify the relationship between categories, and outline the consequences of the processes" (2011, p.364). The process applied in determining final report card percentage grades in Ontario is clearly problematic as it requires the actors to conceptualize student achievement using detailed provincial guidelines, local regulations, as well as personal discretion. Further complicating

matters is a series of real-world issues that should not influence evaluation decisions, but do. Fortunately, CGT allowed this complicated process to be categorized and analyzed to better understand the intended and unintended consequences when completing the task of determining report card grades (Wasserman, Clair and Wilson, 2009).

However, the world does not fit into neatly organized patterns, and the researcher should not try to force data into categories. The result would be a misrepresentation of entire social processes as ultimately rational and consistent. Instead, the researcher can gather the views of participants on processes to help form a validated conceptualized category. Diverging actions and opinions do not destabilize a category, but creates a range of variance that more accurately explains what is happening in a social world (Clarke, 2005). This approach allowed the categories to include comments from all participants with no negative cases. On a descriptive level, the participants' words can conflict. However, on a conceptual level, we see how the situation is comprised of differences, e.g., the co-existence of shared and shadowed practices. To appreciate and articulate the cause, course, and solution of arenas in a social world is to comprehend the situation as a whole.

Although this study took a critical view at the evaluation practices of the participants, it is not meant to denigrate Ontario policy. Those familiar with educational assessment know the complexities of judging student achievement. Large, multicultural education systems like Ontario's need flexibility in policy and practice. Consequently, there will be different translations, and, therefore,

different application. This study was interested in knowing more about these differences. By explaining the process of how participants made their evaluation decisions, we can better understand the praxis of professional judgement in the province. The sample may be limited, but it is detailed, and indicates serious concerns with the reliability and validity of final report card grades in Ontario. A paradigm shift, which is possible in the current framework of *Growing Success*, could bring about the decades-long goal of true assessment and evaluation reform and increase the likelihood of improving student learning through informed professional judgement.

Finally, it should again be acknowledged that the researcher is part of a study's construction. As an Ontario educator, I was looking at an overview of the situation, and, from within, simultaneously. By contemplating data through the lens of one's own experience, the aperture of understanding can be increased – but one must be wary to establish one's own context. Personal interests can help guide the research to areas that require clarification, but it is the voices of the participants that should delineate the research process, thus limiting the likelihood of allowing prior assumptions to influence the findings. If the claims are truly grounded in the data, various Ontario educators could read this chapter and agree and disagree with the views of individual participants – but will confirm the credibility of the claim that in order to complete the job of filling out the report card an educator needs stable ground. This study refers to this ground as Heuristic Assessment.

3.2 Administrative Guidelines

A good place to start our discussion would be the relationship between the local administration and the participants. This category looked at Administrative Guidelines as a conceptualization of what participants *think* they *should* do in regards to evaluation guidelines. By its definition, in order to apply professional judgement, an educator needs to be aware of the conventions of assessment. Since the local administration interprets policy and establishes protocol, a culture that has a tremendous impact on how the participants develop their professional judgement is established. When it comes to how to evaluate Ontario students, as Larry said, "it all depends on the administration you're working for." Since most of the participants have worked under different administrations, they showed how informed professional judgement included the ability to adapt to a different set of caveats. Although there was a range of variance on how administrations explained assessment, especially in regards to shared and shadowed practices, there are common trends. For example, we can see the extent of legacy assessment practices (e.g., calculating grades, normative referencing, and mixed with *Growing Success* directives. Also, for most of the participants, they were accepting of the limits placed on them. Ultimately, both the participants and their administrations had the same goal: to see students succeed. Unfortunately, success is often expressed as inflated grades as opposed to actual learning. The local rules provided a framework for participants to develop a corresponding set of rules to increase the chances of student success. However, instead of allowing informed professional judgement to be active, Administrative Guidelines tend to push

professional judgement in regards to evaluation to an after-the-fact action. Instead of allowing true reflection of a student's achievement, one of the rules of Heuristic Assessment is to consider what the administration desires in regards to student results.

3.2.1 "Self-taught"

Most of the respondents described themselves as "self-taught" in regards to accomplishing the task of filling out a report card. None of the participants in this study recalled receiving any detailed instruction while in Teachers' College on anything regarding the provincial report card. Participants report their curriculum mainly focused on issues such as classroom management. As Sally explained, learning about assessment in Teachers' College, "was something that was lacking a lot… even on my practicums, very few of the teachers would speak to me about it." When the issue was brought up, participants were simply told, "Just use your professional judgement." However, this directive is easier said than done. It is one matter to instruct a new teacher to look for a student's "most consistent and more recent achievement," and another to comprehend what that means.

Most respondents learned their assessment practices by informally conferring with other educators. Catherine was the only participant to make reference to the New Teacher Induction Program (NTIP) that was introduced in 2006. One of the objectives of the NTIP is to help new teachers understand proper assessment practices from assigned mentors (Ontario, 2010b). However, the mentor is expected to cover various topics, essentially providing information on all

aspects of being an educator in Ontario. In other words, how to control a classroom and cope with the teaching profession takes significant time to learn, and how to actually evaluate students becomes something of an afterthought. Consequently, it was not surprising that the participants recalled needing additional support from colleagues. From the beginning of their careers, the participants took an eclectic approach to learning evaluation and deriving the practices of colleagues, thus shaping their incipient professional judgement.

The teaching careers of almost all the participants began after the implementation of the New Curriculum. However, most received instructions from colleagues who taught prior to 1999-2003. Consequently, older practices were noted in the voices when discussing professional judgement. Participants such as Miguel and Smitty talked about their adjustments to the New Curriculum, with the former expressing continued "confusion" with it, while the latter expressed a high degree of confidence and approval of the new model. That said, it does help she received extensive training on the New Curriculum as a consultant for the OME. Since the change is still relatively recent, the noticeable differences in how local administrations approach assessment were understandable.

Since the participants mostly described their assessment and evaluation as self-taught, while also indicating that they received coaching from colleagues and the local administration, there was evidence that these participants had absorbed guidance from their school culture. Furthermore, hearing respondents discuss their experiences at various schools, it was clear evaluation practices largely depended on the school. If Ontario had a truly universal assessment policy, there should not

be such noticeable differences between schools. Instead, participants learn from the immediate environment, and use the information to guide their assessment practices. In turn, these guidelines get embedded into informed professional judgement and become Heuristic Assessment.

3.2.2 "But the computer said…"

Even before the New Curriculum, the use of computer software to help calculate grades was common and encouraged by the OME (Ontario, 1999a). Smitty indicated that in the days of normative-referencing, the computer offered to make the task of assessment easier and, ostensibly, more objective. Even though *Growing Success* made no reference to computer-assisted grades, the use of evaluation software, such as the popular program *Markbook,* is dominant in Ontario schools. Thus, the first legacy issue, and most important non-human factor in the analysis, discussed in this analysis is the dominance of assessment software and its relationship to professional judgement when determining report card grades.

Smitty best illustrated this atavistic practice. She conveyed there was an obdurate feeling among educators that the computer had be obeyed. While consulting for the OME, she felt resistance from teachers when she advised them to use "professional judgement" to determine grades. Moreover, there were misperceptions regarding the use of content standards when evaluating. Smitty's story about her own daughter (pseudonym Jessica) showed how a teacher in 2002

did not feel she could use her professional judgement to determine a report card grade:

> "And this was at the time, professional judgement was being brought in. But the teachers weren't used to it. They had ten years, or more, of just computer, computer, computer. So I said to the teacher, so what is [Jessica's] achievement? And she said, she's getting a 92, but that's not really what she earned. She's more than that. She's more like a 97. So I said, so will she be getting a 97? No, she said, the teacher. The computer said 92. So I said, but in your professional judgement, she's more of a 97? Absolutely. Yes. But the computer said 92. And I said, even though your professional judgement, I said it three times, even though your professional judgement said 97, she gave Jessica a 92. And [Jessica] lost a scholarship because of that." (p.2, L17-24)

The importance of computer-generated grades to administrations was still prevalent over a decade later. Twelve of the participants reported that their administrations expected a strong correlation between the computer-generated grade and the mark on the report card. As Lorrie said, "whatever *Markbook* says it is, we go with that." Another six of the participants indicated that they can stray a few percentage points from the computer, but must have valid justification. Only five participants said that software could be a useful tool, but their professional judgement was trusted and rarely questioned. Lucy was the only participant who worked at a school that did not use any software, and she had full discretion over all assessment and evaluation decisions.

Throughout the rest of the categories, passing references to "the computer" are made as it is central to determining report grades at almost all Ontario schools.

Although informed professional judgement should allow an educator to use conversations, observations, and student products to determine a grade based on consistency, most of the participants were under scrutiny from students, parents/guardians, and/or the local administration to ensure objective-only means of determining grades. As a result, computer-generated grading, encouraged by most administrations, placed limits on the extent to which participants could apply their professional judgement.

3.2.3 "Just old-fashioned average"

Programs such as *Markbook* can be set up in multiple ways and are capable of performing different forms of calculation, such as the blended median or blended mode. However, using the average was the most dominant calculation method, with over half the respondents stating it was the only one permitted by local guidelines, or as Catherine said, "just old-fashioned average." Furthermore, most were not advised to parse individual assignments for outliers; marks that were not consistent with other assessments. When discussing the issue of using averages with Lorrie, she stated, "We don't do any adjusting. You're referring to most consistent, that kind of stuff? Like most recent, most consistent? I know other schools are a little more in line with, sort of, taking a student's most consistent mark and using that, but we don't do that." Even in the follow-up questions years after the initial interviews, participants confirmed little had changed in regards to using average to determine final grades.

Reliance on a computer and the use of average were both shared practices from the days of normative referencing (see Section 3.5.6). There was a practical explanation for this legacy. First, the socially-created sense of consistency: if all

teachers used the same calculation method, it created the putative belief that grades became more reliable and valid. Lorrie believed her principal was strict about uniformity at her school of over 1400 students because, in Lorrie's words, "consistency appeases the parents and the kids in the community so nobody's arguing, this teacher's doing that, and that teacher's doing that." There was the appearance of confidence in entering empirical information into a computer. For many students and parents, seeing the mark on a screen, combined with the familiar word "average," as opposed to blended median or blended mode, created the greater sense of a tenable evaluation.

The combination of reliance on the computer and the continued supremacy of using average, further helped to demonstrate how the local administration set the rules for assessment. In turn, shared practices became part of professional judgement, even if the spirit of such directives were the opposite of what individual teacher discretion should look like. New educators come into this environment, and follow suit. This apparent eviscerating of informed professional judgement was further probed to see how it made the participants feel about their assessment practices.

3.2.4 "There's bumping it, and then there's using your professional judgement"

For many of the participants, professional judgement is just a term for "adjusting grades," sometimes at the behest of their administrations. Some participants reported their administrators gave explicit instructions to adjust grades, thus limiting individual discretion. For example, Miguel shared an email from an administrator to staff, which stated:

"for reporting purposes, there is little statistical or meaningful difference between two or three percentage points. In order to save a lot of time scrutinizing and re-visiting course marks… I would certainly ask that you consider rounding up for the benefit of students when the circumstances warrant. Naturally, students should basically get the marks they have earned through hard work and effort and the consequences if they haven't put in the necessary work or effort for success." (Personal Correspondence)

Likewise, at Lisa's school "They have goals for class averages and they kind of want us to meet that and sometimes that may, influence us, to give certain grades that might not be warranted." In other words, professional judgement can be corralled by the local administration. Grades could even be unilaterally altered by an administration. For instance, Denise claimed, "I have noticed that grades are sometimes adjusted without my consent, making me feel as though my professional judgment has been undermined." Jordan spoke of his experience with three different schools, each with its own character. At one school, any student who finished with at least 75% would get upgraded to 80% whether the teacher felt like the increase was deserved or not. This alteration made him uncomfortable, but protesting was pointless as it was local policy. Similarly, both Eileen and Winnie lamented they had worked for administrations where they "had no professional judgement," because ultimately it was the administration who dictated grades.

It should be noted that Eileen taught for the same board as Lorrie, but there is a noticeable difference between the schools. Whereas Lorrie reported specific orders not to change a computer-generated grade, at Eileen's school, teachers were

directed to "bump" or "round up" grades. Individual students who required an increased grade were frequently identified by her administration "because this student needs to get into this program, or this scholarship, or whatever. We need to give them a higher mark, so let's bump it." Eileen made it clear, "there's bumping it, and then there's using your professional judgement to determine a fair grade. And, it was never the latter. We would just bump marks." In other words, the basis of Eileen's claim about having no professional judgement was based on the perception that the administration had taken it away by directly controlling final grades. In order to "have" professional judgment, an educator needs to have control over the assessment process by actually using evidence of achievement to make decisions, i.e., applying informed professional judgement.

At the time of his interview, Oliver reported a similar case of administrative control. Parents had informed his administration that their son needed a minimum 60% average to take a particular post-secondary pathway. The parents were assured the teachers would work with the student in order to get him to that level, and in turn informed Oliver. However, he said, "I have taught this student before, and I know this student, honestly, does not care." He was willing to work with the student, but was preparing himself to fight back if the student did not earn the grade. Oliver was asked what happened in a follow-up question. Sure enough, the student's achievement fell short of 60%, but he decided it was not worth the fight against the administration and the promise made to the parents, so he assigned the "requested" grade. In other words, the administration's guarantee

to the parents circumvented his informed professional judgement(see Section 3.5.3).

Likewise, Greg claimed he worked for an administration that pressured significant adjustments of ten percent and more. He came to the acrimonious conclusion that "professional judgement" is actually a term used "to make people feel guilty." When asked to elaborate, he said "they're almost asking you to bump their marks. I think it's meant to shroud success. We feel that students should be more successful, so we should bump their marks." At Greg's next school, he was surprised when the semester ended with "a promotion meeting." It was a meeting prior to graduation where the teaching staff was given a list of potential graduates. First, they look at any student who was not going to graduate due to failing one or more Grade 12 courses. Meanwhile, the student's electronic file, including a profile photo, was projected onto a screen. The teacher of the course was then asked to either explain to everyone why the student failed. The failure had to be justified to the entire staff. Greg felt that the social pressure of having all teachers and administrations present, in addition to the student's profile looming on the screen, created a sense of culpability. As a result, most teachers increased the mark to 50% prior to the meeting to avoid being singled out. Even when a teacher decided to change the mark at the meeting, only certain answers were adequate. Simply stating the mark was being changed "because he or she needs it to graduate," was not sufficient. An answer such as "he did better on the final exam than on the course work" was acceptable. Knowing what answers were permitted

reinforces the concept of administrative guidelines constructing workplace cultural rules for evaluation practices.

Greg's promotion meeting also looked at all the graduating students who had an overall Grade 12 average ending in a four or nine (e.g., 84% or 79%). Final grades from the previous semester were locked in, but current teachers were invited to increase the grade in order to make the average end in a five or zero – hence making post-secondary applications for acceptance and/or scholarships more competitive. In regards to justification, the teacher was again expected to use certain phrases before the mark can be adjusted. Responses like "standard deviation" could not be used because of Ontario's criterion-referenced system. On the other hand, saying that "I could have assigned a higher mark on their last project" was accepted as it meant giving special consideration to more recent achievement, even though the course mark had already been assigned.

Other participants were asked if they had experienced anything similar to Greg. Murray confirmed his school did something comparable. The others reported a situation more like Denise and Eileen's: there is no formal staff meeting; the administration acted unilaterally or asked teachers for more information on how a final grade was delineated. Many of the participants indicated they worked in schools where a final mark could not end with a "nine;" it must be changed to an "eight or a zero" i.e., a 79% to a 78% or 80%. However, the rule was likely to state that teachers must round up. Administration encouragement by the administration to round up grades is an example of a shadowed practice because adjustments are not discussed with students and/or

parents/guardians. After all, the word "adjustment" is not found anywhere in *Growing Success*, so the procedure would be difficult to transparently justify.

However, the respondents demonstrated a range of variance in regards to administrative control over grade adjustments. For instance, Derek was abhorred to hear of administrations modifying final grades. He said, "that has never happened at my school. I've never even heard of that taking place. I don't think so. I would be very surprised if that ever happened at my school. I don't think that takes place at all." Similarly, Harry stated, "I can't think of any instance where the grade was changed at the end by an administrator, although I have heard of it happening." Likewise, Corey could not recall a time where he felt pressured by an administration to change a student's grade. Nevertheless, these responses also highlighted the idea that the tone is set by the local administration.

What is important here is that all of the participants looked at the control the administration had over teacher discretion differently, thus affecting their perception of the conceptualization of professional judgement. The cases of Eileen, Greg, and Oliver demonstrate that some teachers see the administration as a force that neutralized professional judgement and took over the grading process. Others such as Derek, Harry, and Corey felt a greater sense of freedom over evaluation. At the same time, there are general directives such as rounding up grades that threaten the sense of professional judgement to some participants, but not others. The common thread was that either the participant could acquiesce to the wishes of the administration or find other employment, as was the case with Eileen.

3.2.5 "You really have to mess up to fail"

There was a principle on which all participants concurred: Ontario Secondary School courses are designed to be passed. The culture of passing in Ontario is a blend of shared and shadowed practices. As Corey put it tongue-and-cheek, there are some students who think they have "a get out of fail free card" because they could look around and see almost all their peers passing. Oscar reported, "you really have to mess up to fail. You really have to not come to class." All the participants described another legacy procedure not mentioned in *Growing Success*: if a student finished a course (including the course work and final evaluation) with a grade of 47-49% (in some schools, 46-49%), the grade must be changed. The directions from the local administration was usually worded along the lines that either the grade must be lowered to 46% (or 45%) to indicate, based on the teacher's professional judgement, that the student has not demonstrated an even limited consistent achievement of the curriculum expectations. Alternatively, the grade could be raised to 50% so the credit could be granted. Furthermore, the directions continued that if a student finished with an "earned" grade of 50%, the grade is to be increased to 51% as a "signal" to other educators the student was not given a "gift pass." None of the participants recalled ever lowering the grade in such circumstances; they have always raised the grade. This is another circumstance where an administration could reduce the thinking process behind informed professional judgement to a matter of procedure.

Although some participants objected to the idea of the administration independently passing a student who should fail, or pressuring an educator into

giving a significant mark increase for a pass, no one expressed an objection to these directions about passing borderline students. The general feeling was if a student is that close to passing, they might as well be granted the credit. Helen felts that part of applying professional judgement was to be "open minded" and self-assured with decisions. Therefore, unless a participant was absolutely sure a student should fail, then the student should pass. Whether or not this position was the result of indoctrination by administrations, for participants like Helen, it seemed like they were the ones making a fair and equitable decision, thus applying informed professional judgement (see Section 3.3.4). However, automatic actions appear to have less to do with reflection and more about following a personalized approach to evaluation. That said, passing a student is simply easier than justifying a failure.

It was stated earlier that Eileen and Winnie both said they have felt they had "no professional judgement." Eileen left that school as a result. Winnie spoke about the difficulty she encountered when changing schools. Her first placement was a temporary contract where she felt "defended" by her administration. She then spent several years with another school before moving on again. One of the major reasons why she left was because of the pressure she felt to pass students who had not earned the credit. She explained the administration was,

> "not berating you, but there's almost, they're questioning you, why did you give that mark? Why are they not passing? What were they doing, what were you doing? And the thing is, if that student wasn't deserving of the mark, I don't feel that I should have to defend it, you know, look through their work. I always keep

samples of students, especially the students that are failing." (p.7, L16-20)

Because most students are expected to pass, only a few participants stated that before boosting a mark for a borderline student, a meeting with the department head was required. The others said they could simply round up the mark to 50% without any discussion. Participants reported they have never been questioned about a passing grade.

On the other hand, if a student failed any Secondary School course, an explanation to the administration was always required. In cases where the reason for the failure was obvious, most participants did not feel their professional judgement was being questioned. For example, Denise described the first time she had to fail a student: "He didn't hand in enough assignments… so he got zeros on those and never came to class. Even the assignments that he eventually handed in were not up to standard in any way." Furthermore, failure should not come as a surprise to anyone. Participants confirmed it was vital to document presages and actions taken, such as discussing concerns with the student, notifying parents/guardians, and talking with department heads. As a result, the teacher had already built a defense to justify the failing grade.

In practice, most of the participants indicated that if a student finishes with a grade of 40-49%, they are most likely to increase the grade to 50%. Participants summarized this general trend as a means to avoid having to have to explain to the administration and/or parent why a student failed, and possibly have to defend

179

what actions were taken, as in Winnie's case. Hence, professional judgement guided educators towards pragmatism. Derek explained,

> "I think you have to justify why you are failing that student. The onus is on you to say why that student didn't pass… there's less justification needed for why you didn't pass the student than why you passed the student. Your higher ups want to see those passing percentages at all levels. And everyone is trying to get a better pass percentage." (p.7, L5-9)

When it comes to passing, we see an aspect that was perhaps the closest to a universal rule in Ontario schools. Administrations want students to pass, and so did the participants. It was one thing to feel pressured to increase an already passing grade, but there was less opposition to directives of passing borderline students. Part of learning the local school culture included knowing what battles were worth fighting, which became an integral part of one's Heuristic Assessment.

3.2.6 "I've generally agreed with the limits that were expressed by administration"

Despite the range of variance in how the participants viewed administrative control, the desire to be in good standing with superiors could be noted in the voices of the participants. Even those who expressed objections to some administrative guidelines would rather be in a situation of mutual understanding. In order to achieve agreement, Oscar said it is less frustrating to "agree to disagree" instead of "fighting the same battles over and over." The exigencies of teaching lead to the desire of have the backing of the administration. In order to get this support, it could mean conforming to the perception of how the administration wanted to see professional judgement applied, thus modifying one's

professional judgement. As a result, as Lorrie stated, "I don't know if I'm been

brainwashed, but [the limitations are] fine with me." Most of the participants

came to agree with the barriers and have incorporated them into their professional

judgement. Harry said, "I've generally agreed with the limits that were expressed

by administration… I've always been happy with the marks I've submitted."

Similarly, Derek claimed, "At my school, I don't feel the need to have to justify

my decisions." Larry reinforced this notion with the comment "I've never worked

for an administration that has had a problem with my professional judgement. As

long as my school gives me the freedom to use professional judgement, gives me

some leeway… then I don't see any huge problem" with following its guidelines.

When asked to elaborate, he replied,

> "I feel totally free, because I understand the context in which
> professional judgement is given. I know I can change marks, if I
> want, I know what to focus on, I know at the end of the day it's
> fairly subjective to put higher weights on when I'm marking
> assignments, but I also know what the expectations of the
> administration are, so if I can balance those two things, so I can use
> my professional judgement and no one in the administration is
> going to care because they still get what they want and I'm doing
> what I think is best. So if I balance those two, there's no problem."
> (p.10, L18-24)

Larry's use of the word "context" was telling. It suggested that professional

judgement was a combination of assessment knowledge and thinking about the

overall situation; how his decision reflected the desires of the administration.

Professional judgement should be a matter of individual discretion, but it was

often more of a case of following Administrative Guidelines. As Larry suggested, his role is to "balance" these two sides in order to be seen as an effective assessor.

However, it was noted some participants have left schools instead of conforming. In justifying her stance against her former administration, Eileen said, "I think we're best when we're given some freedom to exercise our own ideas." Eileen admitted, "I think consistency is important, to an extent. Absolutely. Things have to be fair… [but] I think that some inconsistency or diversity is good for the student." Again, professional judgement was about having a sense of "freedom" over making decisions. At the same time, the fundamental principle of being "fair" was present. It was not that some of the participants wanted to be able to do whatever they wanted on the report card. They simply aspired to arrive at conclusions on their own while respecting the rights of students, parents, and other stakeholders.

The "balance" between accepting controls, while maintaining a sense of freedom, within the conceptualization of administrative guidelines, appeared to be in the ability to explain how a grade was determined. According to Jerry, "I think I am free to use [professional judgement]… whatever decision I came to, I'd better be able to communicate it to the parents, the student, the principal… if I do that, I think I will have the support." Likewise, James feels, "I have all the freedom in the world. When it all comes down to it… I think it's important that we were able to… justify a mark." George, speaking as someone who was making the transition to becoming a department head explained a teacher should feel "If [the administration] wanted to come in and look [at my marks], they were completely welcome." As a department head, he felt that "trust" between the administration

and teachers is needed for an effective professional judgement. Even in cases such as Greg's, although having pejorative feelings towards the situation, he is not against the need for administrative guidelines; it is the manner in which they are presented that has bothered him in the past.

3.2.7 Administrative Guidelines and Professional Judgement Summary

As illustrated on Chart 3.1: Professional Judgement and Administrative Guidelines, this category was a conceptualization of how the participants perceive the guidelines they are given by the administration, which in turn influences how they understand professional judgement. Overall, participants wanted to be compliant with administrative directions, and generally agreed with instructions. However, administrative oversight could be the antithesis to professional judgement. If the latter was about personal discretion, the former was about enforcing consistency. Therefore, for some participants, "professional judgment" was a set of procedures. These procedures comprised entering marks into a computer, looking at the calculation(s), then making student-friendly adjustments. Adjustments included passing weak students, because it was widely believed passing a student met with less resistance than failing one. Since professional judgement needed to account for factors such as "context" and "criteria," following guidelines as part of professional judgement was not paradoxical (Ontario, 2011a, p.152). Much depended on the culture of the school, as constructed by the administration, because this is turn established the conceptualization of professional judgement. Participants absorbed this culture,

and demonstrated it in ways such as their assessment practices. Some

administrations encouraged the thinking aspect of informed professional

Chart 3.1: Professional Judgement and Administrative Guidelines

Evaluation Practices
were "self-taught"
via school culture

Most of the participants
were instructed to go by
"what the computer said"

**Administrative
Guidelines**

"it all depends on the
administration you're
working for"

Most consistent / more
recent achievement was
usually expressed as "just
old-fashioned average"

Participants "generally
agree with the limits
that were expressed by
the administration"

Administrations encouraged
participants to find ways to
pass students: "You really
have to mess up to fail"

Professional judgement was
a "balance" of a participant's
"freedom" to evaluate based
on assessment knowledge,
within the framework of
local rules and a tendency to
raise grades

Administrations often
encouraged positive mark
adjustments: "There's
bumping it, and then there's
using your professional
judgement"

judgement and gave educators more discretionary power, allowing participants to

feel guidelines were properly followed while their own autonomy determined the

final grade. Generally speaking, respondents recognized they were in a profession

where their decisions can have a major impact on the lives of adolescents, so some

degree of oversight was needed to ensure evaluation was fair, transparent, and

equitable for students, even if some practices were done in the shadows. Part of

professional judgement was the ability to process competing factors and achieve

balance. However, the way administrative guidelines direct student assessment,

even when done in the interest of student success, raises reliability and validity

issues (e.g., automatic raising of grades and the suggestion of considering effort).
The next category is also an area where teachers coped with dissonance within
another area of administrative guidelines: the application of punitive measures.

3.3 Punitive Measures

The reinstatement of punitive measures was a surprising policy turnaround
in *Growing Success*. Although late work, missing work, and academic dishonesty
are widely considered to be under the umbrella of student behaviour, and student
behaviours are not to be considered as part of determining a final grade, policy
made an exception in this regard. Ontario teachers may consider the use of
punitive measures on individual assignments (i.e., late marks or the use of zero), as
long as it does not distort true overall achievement on the final grade. Therefore,
professional judgement is needed to sort out these instructions. Harry claimed
having punitive measures "does make a difference in encouraging students to
complete their work and to complete it on time." How the participants applied
punitive measures and adjusted final grades accordingly was a matter for
administrative guidelines as well as personal discretion. Different schools had
their own rules and instructed participants accordingly. The conceptualization of
what punitive measures are and what they were perceived to do when recording
student achievement is a significant part of understanding the application of
professional judgement. Although the participants supported the option of
punitive measures because of what they see as life lessons, they did their best to
shelter students from having the report card grade dragged down.

3.3.1 "In the real world you'll get severely reprimanded"

With the return of punitive measures, the participants felt that late/missing assignments and academic dishonesty have decreased. At least that is how they interpreted the situation. However, late work and academic dishonesty were still issues for at least one student in every class. Participants tended to ask the same question when counseling such students: "what are you going to do when you get a job, and if you do not do it properly – you may get fired?" Thus the basis of the participant's conceptualization of punitive measures is that they provide an analogy for what happens in the "real world" when responsibilities are neglected.

When participants spoke about punitive measures, without prompting, almost all alluded to a workplace analogy. Elaborating on his support for punitive measures, Harry said, "it has an impact on them performing at later stages in their life be it post-secondary or with a job in the future." Likewise, Oscar felt, "We as teachers would not be doing our job if we did not prepare students for the reality of life. One of those realities, unfortunately, is that you have to hand in your work on time when you are in a professional setting." Reflecting on an educational system without punitive measures, Oliver claimed,

> "essentially you're telling an entire generation of kids to think they cannot fail. And once they graduate from school and get out to work, you can't have that attitude saying oh well, I know this project was due today, I'll hand it in next week. The repercussions for that in the real world or you'll get severely reprimanded." (p.2, L33-36)

Again, a student's report card grade should only be an expression of overall achievement. However, what is the place of punitive measures? Without punitive measures, according to the participants, students develop poor time management skills. Although policy permits punitive measures, they should not interfere with a student's true level of achievement. The participants essentially agreed with this position, however, how they graded students with punitive measures was an example of Heuristic Assessment as opposed to informed professional judgement due to their own approach to evaluation and the rules imposed by the local administration.

3.3.2 "Avoiding failure"

We must be careful to not go off on a tangent and discuss how punitive measures are applied to individual assignments. The focus of this study is on the application of professional judgement to final grades. For instance, when a participant is deciding on a final grade, and punitive measures have been applied, what happens? First, we have already determined Ontario's preference for passing grades. Participants had no issue with passing borderline students. More often than not, the reason why a student is borderline is because there are assignments with punitive measures applied, including zeros. However, the local administration may request that the student passes. A 50% becomes a "gift pass," a common phrase among the respondents. All participants were concerned that gift passes meant no life lesson for the student. However, the emphasis in Ontario Secondary Schools is on "avoiding failure." Participants acquiesce to administrations if a gift pass is requested. It is not the respondents were upset

about any student passing; there are too many other concerns to dwell on it. Still, gift passes made the participants worry about the student's future success. Furthermore, gift passes added to the feeling that informed professional judgement is being taken away, and evaluation became Heuristic Assessment.

As previously discussed, it was usually the participant who unilaterally decided to pass a borderline student, thus most likely limiting the impact of applied punitive measures. This decision is made with the knowledge that it is unlikely that the student will achieve dramatically better in the future. Furthermore, if there is a history of academic dishonesty and/or late/missing assignments, this behaviour is likely to continue in cases of borderline students. Murray commented that he felt students "get about the same mark in every course," despite the suggestion that if assessment should improve student learning, grades should increase. Respondents admitted that "passing the student along" is an issue. At the same time, there is the reality of having to pass students in most cases. It seems it is better for the student to attempt the next stage, then to hold them back and try again. This is the dominant message by the participants, via their local administrations.

Similar to how schools varied on to what degree administrations adjust final grades, directives on punitive measures differed greatly from school to school. Some left the decision to the individual teacher, some targeted a single learning category to be deducted, some indicated that the entire student product be penalized. Punitive measures ranged from the deduction of a sub-level (e.g., a level 3 to a 3-) as long as the assignment was handed in, entire levels per day, to an

automatic zero after four days (see Section 3.5.1). Again, an entire study could be conducted on punitive measures as each participant provided a different situated description. In the interests of the present study, it can be said that students of the participants would not fail a course if late marks alone brought the computer-generated average to below 50%. Missing assignments and consistent academic dishonesty could result in failure, but again, it is preferred to avoid issuing a failing grade.

3.3.3 "It's a definite professional judgement case"

This study noted many contradictions between policy and practice, but with punitive measures there was more consistency. None of the participants reported that their administrations prohibited punitive measures. Half of the participants claimed it was requested that punitive measures not be used, and saved as a "last measure" and on a "case-by-case" basis. For half the participants, the application of punitive measures is a matter for professional judgement. The general feeling is the measures are more effective with academically-inclined students who do not want penalties to interfere with post-secondary applications. Not that the participants want to apply punitive measures, but for students who want to attend post-secondary studies, it is felt the presence of punitive measures "scares students straight." When punitive measures have been applied, and the student is still not in danger of failing, what happens to the final grade? Again, policy stated that the application of punitive measures should not distort overall achievement. Although the participants find this statement agreeable, we have already seen that some participants have little room to maneuver when

determining a final grade – even if a student demonstrated evidence of learning on assignments that did not receive punitive measures.

Eileen, who has made it clear how important evaluation discretion is to her, explained she accounts for punitive measures on a final grade by "considering the student's performance over the whole semester, as well as any important extenuating circumstances that may have contributed to late or missed assignments," and determines a grade according. Corey's position on punitive measures was, "it's important that there's some freedom for the teacher or the administrator to come up with a decision that fits that student and is going to best serve that student." The first thing a teacher should ask, according to Dirk, is "Why did they miss it? There's going to be different circumstances… I would say it's a definite professional judgement case." After all, acting in the best interest of the student was the immanent quality of informed professional judgement.

A major consideration for the participants was the age and maturity level of the student, a position that was reflected in *Growing Success*. Many students do not understand the concept of "the real world." As Murray said, "I'm all in favour of giving chances, because, if you're 14, 15, 16, 17, you'll make mistakes. You know? You're not fully aware of how the real world works." Through counseling and multiple chances, it was hoped younger students will develop their learning skills and work habits to establish better time management and gradually show improvement. Likewise, Lorrie was more lenient with Grade 9 and 10 students regarding missing assignments. She had them do some work in class for marks because once a grade was entered it could not be changed. However, she had less

sympathy for Grade 11 and 12 students, even though punitive measures would automatically impact the report card grade. She hoped by that age, and with intentions of postsecondary study, students should know how to meet deadlines.

However, half of the participants reported some level of mandatory punitive measures they were supposed to enforce. For example, Jordan and Catherine must assign a mark of zero if an assignment is not handed in within four days of the due date. Combined with using average and needed to report the grade from the evaluation software, one zero can significantly change a student's final mark. Catherine and Jordan would prefer if their administrations allowed them to base a mark on median or mode so these individual penalties were less harsh. Furthermore, they would prefer to have the option of whether or not to apply punitive measures in the first place, but it is expected to be a shared practice so that the measures are applied evenly. Although the participants universally agreed that the option was necessary, overall, they did not like applying mandatory punitive measures as every student is a case-by-case basis. The situation is already undesirable when participants need to consider negative marks. Either the student has not handed in an assignment, or has tried to cheat. Having to address such issues, including getting parents involved, was not an enjoyable part of the profession.

3.3.4 "You know your students"

Again, there was a range of variance when it comes to the application of punitive measures. Participants approve of the option, but then there is the matter

of applying them fairly. When administrations tried to force their use, they could take away from an educator's professional judgement. At the same time, many participants felt that some students needed a life lesson, but did not want to take away from a student's true level of achievement. Consequently, the participants used Heuristic Assessment to engage in shadowed practices. For example, Helen told her students they would get punitive measures for late or missing assignments, but only applied them to missed presentations as they disrupted the learning process for everyone. Alternatively, if one of Miguel's students had more than one late or missing assignment, he would try telling the student that marks had been deducted, but a higher mark was entered into the computer. Alternatively, other respondents, such as Larry, Lisa, and Winnie had a strategy to limit the impact of punitive measures on assignments throughout a course. They went back and looked at all assignments. If a student had since shown better achievement of the same curriculum expectations where marks were lost to punitive measures, they changed the entry. In other words, they removed the punitive measure because it is no longer seemed justifiable. Winnie explained that she had to use "average. And even sometimes, when the average is calculated, I might, again, use my professional judgment and change that to how I see things in my classes. And because at that point you know your students."

Winnie raised an important point: the final grade should reflect what an educator knows about a student at that point in time. Past punitive measures, or any marks that do not reflect the consistent level of achievement, based on *Growing Success*, should be rectified. However, to do this transparently was

difficult because stakeholders, according to the participants, expected grades to be based on objective numbers and not teacher discretion. To modify individual evaluations from over the course can come across as teacher bias, even though the intention is to update the overall record to reflect true achievement. Still, to trust professional judgement to this degree, despite its prominence in policy, appears to be a tall order.

On another note, there may be compassionate reasons to not apply punitive measures, depending of the student's circumstances. In this regard, all participants would fight mandatory punitive measures if a student suffered a known personal tragedy, and it is likely all local administrations would make an exception (see Sections 3.4.1 and 3.5). Therefore, to identify what qualifies as a "tragedy" requires professional judgement. As we will see in upcoming categories, rather than engaging in a prolonged debate over the most fair, transparent, and equitable course of action, most of the participants simply adjusted the overall grade based on local guidelines and professional judgement.

3.3.5 Punitive Measures and Professional Judgement Summary

As seen in Chart 3.2: Professional Judgement and Punitive Measures, this category was relatively the smallest in the study, but needed to stand out. It has its own range of variance, and is its own arena in the situation. Punitive Measures not only has elements of Administrative Guidelines, but also expanded the notion that participants consider their students as people (e.g., their maturity level) and not

just entities of achievement. Informed professional judgement is used to make

decisions that impact the lives of students. The respondents wanted to prepare

Chart 3.2: Punitive Measures and Professional Judgement

Participants supported punitive measures because "in the real world you'll get severely reprimanded" for not doing work

"Avoiding failure" was protocol, so punitive measures alone would not cause a failing grade

Participants say "you know your students" when determining the final grade, so they try to limit the impact of punitive measures on true achievement

Punitive Measures

"[do] make a difference in encouraging students to complete their work and complete it on time"

When it comes to punitive measures, respondents agree "it's definitely a case of professional judgement"

Participant professional judgement generally supports punitive measures because they serve as a life lesson, but are used as a last resort

their students for the real world; they cared about their students' future success.

This point is important to note, as policy advised to only consider achievement for

evaluation – but to also use assessment to foster life long learning. To foster life

long learning, it is felt that punitive measures can be a "wake-up call" for students.

Punitive measures are seen as an effective deterrent to late/missing work, but the

participants were not interested in punishing students. It is a tactic to try to get

their attention. At the same time, it is recognized that other tactics are needed for

other students, such as those who are in danger of not making it to graduation.

Therefore, participants preferred to have the autonomy over punitive measures. Furthermore, how participants felt about administrative guidelines on punitive measures, and how these measures are sometimes circumvented in practice, revealed key details about professional judgement. Informed professional judgement should allow a participant to act based what is know about a student in order to provide an effective evaluation. When participants felt the administration was interfering with evaluation and practices happen in the shadows, Heuristic Assessment occurred. The next category continues the discussion of the link between other student behaviours and the impact on professional judgement furring the final report card process.

3.4 Perception of Student Behaviour

As previously stated, *Growing Success* advised educators to eschew student behaviour when evaluating the percentage grade. The learning skills and work habits section of the report card may address behavioural issues, along with some parlance comments. However, the participants of this study reported behaviour is extremely difficult to ignore when applying assessment of learning. Indeed, we have seen evidence that some administrators encourage consideration of perceived student effort when grading. Earlier, Oliver referred to a student "not caring" about school, which made effective teaching difficult. It was also pointed out how Lorrie's report card grades were scrutinized. But when asked if she could unilaterally round up a grade, such as 79% to 80%, she replied, "Oh, well, come on Gord. Everyone bumps 1% for those strong students!" Such statements revealed another aspect of the conceptualization of professional judgement: it is

not that the participants were consciously rewarding and punishing students for actual behaviour, but reacting to the perception of behaviour informed by classroom interaction. How the student viewed the learning process, based on the respondent's perception, contributed to Heuristic Assessment and effected grading decisions. Participants did not expect to be treated with constant obeisance, but did want to see students engaged in the learning process. Participants expressed frustration with obstinate students who did not follow course work instructions and/or general school rules, such as attending class. Rebellious acts in themselves did not result in a grade reduction. Recalcitrant students may have acted out due to external factors. On the other hand, students who displayed positive behaviours – such as effort – received approbation from participants in the form of a slight grade increase. It was a shadowed, but quite common practice. These participants suggested reflecting on student behavior is an ineradicable factor of the grading process and Heuristic Assessment.

3.4.1 "I'd like to say behaviour is not a factor, [but] I think that's unrealistic"

Ontario Secondary School teachers spend approximately 110 hours of scheduled instructional time with a student, and there is the possibility of after-school counseling, as well as other opportunities, to augment instruction and gather evidence of learning. Through observations, participants discovered how a student acts in the classroom. Here we seen factors associated with symbolic interactionism and SA: participants saw the evidence, interpreted it, and then internalized it. This process formed an influential perception about how a student felt about his or her education. As Corey said, teachers "take in a lot of different

196

factors when evaluating students. As much as I'd like to say that behaviour is not a factor, I think that's unrealistic." He continued, "As a teacher, I work as hard as I can to avoid having that be a factor." Likewise, when asked about the importance of a student's classroom behaviour, Murray stated: "It's huge! I mean, as they say, attitude is everything, right? And your attitude… is who you are. It's the direction you want to take to success." Again, it is not that the participants were evaluating student attitude itself, but how students behaved made an imprint on participants.

All participants identified what they considered to be positive and negative behaviours. In regards to negative behaviours, Dirk summarized, "there's that student who doesn't do his homework, misses class, comes into class late every day, effort is low… that feeling you have at the end of the semester is going to be negative." When students consistently displayed a lack of interest in learning, it could enervate the participants. As Derek explained, "At the end of the day, a teacher's job is made most difficult by a student who is not willing to meet the teacher half way. A student must be willing to put in an honest effort in order to have success." Derek promulgated, "I have three very simple rules for my students: 1) Come to class 2) Do your work 3) Show Respect. I tell students if they follow those three rules they will pass this class. If they follow them well, their mark will reflect that." Explicitly and implicitly, through shared and shadowed practices, these participants informed students of certain behavioural expectations and the impact they would have on the percentage grade. Learning is a social process, which requires people to act in certain ways to maximize chances

for success. A poor showing by the student imputed a perception of negative behaviour. To reiterate Corey's comment, the professional educator needs to take such feelings in stride and try not to allow them to interfere with evaluation. At the same time, we can see in Derek's words, not following basic behavioural rules could result in a poor grade, including failing a course (see Section 3.5.4).

Participants were asked if they tend to label students based on behaviours. Catherine, among others, said, "I do agree that we sometimes put these labels on students." Being "lazy," "not caring," or being "hard working" were recurring codes. For some participants, caring about a course was necessary for success. Denise said students "do have to care about the course to an extent… students need to understand that it's not just about doing the work, you need to follow the rules." Similarly, Lisa remarked, "you get to know them and who's working hard and who's not working hard." Again, we see the notion of knowing a student, this time in the regard of work ethic, and how it was translated into a percentage grade.

Like Oliver, Derek explained he applied a "not-caring label" for students who missed a lot of classes, and/or was "disrespectful to the teacher or others in the class." He added that he tried to encourage all students to care about their studies because "a good teacher will be able to inspire that effort based on engaging lessons or interpersonal relationships." However, a participant's best efforts to engage students were not always successful. For example, it was impossible to inspire a student who was not in class. Although attendance is not evaluated, it is essential that students come to class in a system where 70% of the grade was based on course work. Oliver elaborated on a key difference in his

History and Drama courses: in the former, a student could catch up through homework. In the latter, however, so much was done in class through interactive lessons, that habitual absences would leave the student far behind. Therefore, students needed to show they cared by simply showing up for class.

These points do not mean that all students fit into either a positive or negative box of perception; here there is another range of variance. As Eileen wrote in a follow-up question, "I usually have a few students that seem to 'not care' about the course. And I think that I am usually right. But I don't see TWO [sic] groups of students in my classes -- those who care and those who don't. I think it is more complicated than that and I have never felt that division." All the participants could think of students who clearly had a negative attitude, but still did well in the course. George stated students in his courses "who act in inappropriate ways usually did poorly in class." However, he had "some students who acted very badly in class… but were intellectually gifted… if they do really well on the test, then they're going to get the mark." In other words, negative student behaviour did not mean a participant would be the subject of bias when the said student demonstrated evidence of learning. Regardless of their general attitude, the participants suggested students who fulfilled the success criteria for the learning goals would get the grade they earned.

Similar to how participants considered the student as a person before applying punitive measures, the perception of a student's behaviour also works its way into determining grades. The local administration may expect teachers to uphold punitive measures for the sake of consistency, while also instructing

teachers to disregard other behaviours. Hence confusion was added to an already

convoluted process. Participants solved this dilemma by doing what they think is

the most fair and equitable course of action. They did not use grades to punish

students with negative behaviours, but those who display positive behaviours are

more likely to get the benefit of the doubt in the form of a favourable evaluation.

3.4.2 "A lot of that has to come back on classroom-management type techniques and understanding the kids"

The cause of poor performance or negative attitude may not be about

academic ability or carelessness on the student's part, but external factors such as

socio-economic background, family issues, personal struggles, etc. Ideally, when

a participant became aware of "outside factors" that were harming student

performance, the school's counseling services could try to assist. If a student's

day-to-day performance could be improved, then it could translate to greater

achievement. Lucy had an excellent anecdote concerning such a student: "For the

first few months he barely showed, hardly handed in anything, and never

participated. Then the school social worker helped him… he was just

overwhelmed with what else was going on in his life, and so school was

understandably not a priority." Lucy's excerpt showed this student could have

been dismissed as "not caring" if action had not been taken. Therefore, part of an

educator's professionalism is to use assessment for learning to uncover the source

of the problem. Smitty pointed out: "Usually the child is acting out because they

don't understand…you don't penalize children, or the students, because of their

bad behaviour, I think that, a lot of that has to come back on classroom-

management type techniques and understanding the kids." Again, because of the

potential amount of interaction between an educator and a student, there were

opportunities to get to know a student. Sometimes knowing a student included

evidence of external factors portending poor achievement. At the same time,

classroom management can impact the behaviour of students, i.e., an educator

must show respect in order to earn respect. However, a teacher cannot work with,

or get to know, a student who does not come to class. The absences, combined

with a lack of knowledge about the student, is usually rationalized as a negative

behaviour. Furthermore, the lack of attendance prevents the teacher from

gathering evidence of learning, adding to overall teacher frustration.

Participants often reflected on themselves when thinking about how to best

help students and understand their behaviour. Some participants admitted they did

not care much for school as an adolescent. Jerry made a good point about

participation, saying "As a student myself, I didn't really participate. Different

students have different learning types… [besides], a student who raises their hand

a lot can also be way off topic." Likewise, Oscar pointed out some students could

be "naturally shy," but "study their brains out outside of class" and do well. He

added, "I think that another factor to be considered is that they are still teenagers

and in general teenagers are distracted with so many things that are available to

them, like *Facebook* and *YouTube*. With so many better things to do why do the

things that are not fun to do?" As Derek alluded to earlier, it is part of an

educator's job to establish a stimulating and inviting learning environment.

In short, poor course performance was often linked to behavioural issues.

Nevertheless, there was no evidence that participants used grades to punish bad

behaviour – it was just difficult to help students who resisted the learning process. Again, it could be noted that students who earned a grade below 50% were still likely to be given a course credit. However, students who demonstrated that they take their education seriously by displaying what was perceived to display "hard working behaviour," were more likely to receive a small increase to their percentage grade. In a participant's professional judgement, it is not favouritism, but the right (i.e., fair and equitable) action to take.

3.4.3 "Go the extra mile"

To show that they take their education seriously, students need to demonstrate a modicum of effort; it is the sine qua non for achievement. Perceived effort was usually rewarded with a positive mark adjustment. Effort comes in many forms, including regular attendance, contributing to class discussions, and asking for clarification on teacher feedback (see Section 3.5.2). Catherine explained, "if a student worked really hard, displayed a lot of effort, and got a 78, I'll likely increase it to an 80. However, if the student is capable of getting a 90, and ends up with a 78, then that's what he gets." Likewise, Miguel said that when he reflected on the software-generated grade, he thought about his "impression" of the student and "if it [the grade] reflects their actual effort." Furthermore, he would not "boost" a student's grade if there had been any attendance issues. Similarly, Lucy stated, "When I see a student with an 87, and I'm thinking, well, is the effort that they've shown, and their initiative, all those other things, have they done enough to merit bumping that up to a 90 or not? That is essentially how it comes into play." The participants viewed such positions as

being more objective than subjective. A participant would not assign a high grade to a student who was visibly trying to do the best work possible – but the products were still poor in quality. However, the effort alone could earn a few extra percentage marks. Participants preferred to see all students trying to do their best. Preference was different than bias because it recognized there is an ideal, but does not admonish those who do not live up to the ideal. Again, it is the talented teacher who can work with students to help them reach their full potential, but there can be no full potential without effort. At the same time, using effort to help determine an overall grade is not the intention of informed professional judgement, but further reflects the practice of Heuristic Assessment.

Participants could not discuss high grades without at least hinting at the need for students to put in effort. This position was evident when participants discussed Level 4 achievement. When explaining the difference between getting 85% and 95% on an assignment, Oscar explained 95% went to those "who really go the extra mile... it is a hard thing to quantify, and again it comes down to professional judgement… You have to impress me. You have to impress the teacher… It's the wow factor." Lorrie said something similar, including the words "go the extra mile" and "work their butts off" when she thought about the ideal student. Participants echoed the desire to develop intrinsic motivation in students. They wanted to see their students take initiative, demonstrate independent learning, and, as Miguel said, to "not need to be spoon fed" or be "chased after" to get work done. To the participants, such behaviour should be rewarded as it is seen as a life skill and congruent with the fundamental principles.

Participation was also a form of effort desired by most participants. Helen pointed out, "if no one participated, it would be a really boring class," so she told her students she would award marks for participation. Jordan said, "I think if the student is very participatory I guess there is some reflection of that in their marks… if I'm waffling of a student between an 82 and an 83, I would think for a second: are they participatory in class? It's an 83 if they are." Other respondents also commented on the importance of participation. As previously discussed, the respondents recognized that there are different forms of participation. A student may be shy, or culturally bound not to share ideas in front of a group (e.g., a common practice among some indigenous students). However, active participation in the learning process could also be observed in smaller groups or in one-on-one consultation.

Participants also reflected on a student's overall effort during the semester, especially those who faltered towards the end. George explained that sometimes good students break down, which is why it is important to remember a student's overall consistency: "If they show they really worked hard, they look broken at the end of the semester… then you would bounce them back up because you realize the pressures of life… just burnt them out." If the products later in the course regressed from earlier evidence of learning, and the participant believed that outside factors played a negative role, professional judgement would also take these assumptions under consideration. However, assignments later in the semester usually had a heavier weight in the spirit of special consideration for more recent achievement. Since most of the participants used average, instead of

allowing professional judgement to determine special consideration, the more recent achievement was automatically factored in the percentage grade. Therefore, in yet another example of shadowed practices, participants could give these individual assignments a higher grade than they deserved, so the computer would generate a grade consistent with what the participant saw as true overall achievement.

Effort could also be conceptualized in the form of improvement. Some of the participants considered a student's overall improvement as part of special consideration for more achievement. For instance, Winnie was "more apt to give [students a mark increase on the report card]… because you know they really did try hard and they really did try to come along from their last essay and they did try and improve." Likewise, James said, "I can see if a student is improving or taking what I've told them in a counseling session and actually applying it. If I feel that they're not making any movement, then I think it reflects a lack of effort on their part." In both cases, the participants feel they were more likely to increase a student's grade on the final report card if they could see the application of their feedback, resulting in overall improvement. On the other hand, they would not penalize a student by lowering an earned percentage grade if the student has not demonstrated any noticeable improvement.

In regards to passing borderline students, although their behaviour was likely a major cause for being borderline – it could also be their saving grace. When a participant was debating whether or not a student had demonstrated a limited achievement of the curriculum expectations, Heuristic Assessment

considered overall behaviour. For example, Eileen said, "were they present in class? …it's different with every student, and a lot of the time it comes down to your professional judgement… It's their attitude, it's their effort, their willingness to learn, it's, have you seen change? Progress over the semester?" Lorrie made a similar comment, stating if there had "been significant improvements with [poor] students and special situations in some of those cases we have been considering most recent or most consistent depending on the scenario." She continued, "But for each individual student in each class (without any issues), we still take the calculated mark." As Helen put it, "My course is pretty straight forward… It's the students who don't care, don't come to class, don't hand in assignments who are at risk [of failing]. As long as they've tried, they'll pass the course." Likewise, Corey stated, "I would fail a student that doesn't submit assignments, there's a lot of missing assignments, high absences, generally somebody who doesn't care. And you know, somebody who just hasn't shown any effort or evidence of learning throughout the course." Again, we see the subtle difference between the Perception of Student Behaviour and Punitive Measures categories. Both are about behaviour, but the latter is permitted to count towards a student's final grade, hence why they were divided into two categories. However, students with late or missing assignments often have other negative behaviours that, when taken together, could result in failing a course. If a student had late/missing assignments, *or* other bad behaviours, in most cases, the credit could be saved.

3.4.4 "Teachers (and everyone else in the world) definitely make judgments based on perceptions!"

If the perception of student behaviour impacted professional judgement when determining grades, can this impact be quantified? Eileen said, "teachers (and everyone else in the world) definitely make judgments based on perceptions. In fact, that is what we are paid to do! So finding a metric to use as an assessment tool in conjunction with our perception is a difficult task." Reflecting on a student's behaviour is both a shadowed and shared practice in the sense the practice was be specifically forbidden by policy, but for the participants, it is unrealistic to disregard. However, it is still a case of Heuristic Assessment. Considering behaviour may seem fair and equitable if not entirely transparent, but it interferes with reliability and validity. Ultimately, as Corey explained, the participants would assign "a grade that will also tell them that they need to work harder or congratulate them, or kind of reward them for the work that they have achieved." Such remarks further revealed the fallible nature of professional judgement, but for the participants the position could be objectively explained under the direction of Heuristic Assessment.

Corey's use of the word "reward" is troublesome. The participants did not specifically suggest they actively assigned bonus marks for good behaviour. From their point of view, whether one wanted to call them "bonus" marks or not, they were simply assigning a slightly higher mark to those who displayed effort. As we will see in the next category, the quantitative aspects of these adjustments became more apparent when compared to the levels on the Ontario achievement chart. Again, the participants demonstrated their own sense of consistency in regards to

being fair and equitable. However, being fully transparent at the same time is

actually the most challenging part of the whole process.

3.4.5 Perception of Student Behaviour and Professional Judgement Summary

Chart 3.3: Professional Judgement and Perception of Student Behaviour

illustrates how this category expanded on the notion that it is extremely difficult

not to consider the student as a person in an internal-assessment system.

Chart 3.3: Perception of Student Behaviour and Professional Judgement

Participants recognized that "a lot of that [student behaviour] has to come back to classroom management techniques and understanding the kids"	Participants knew that they should only consider achievement, but "I'd like to say behaviour is not a factor, [but] I think that's unrealistic"

In order for a student to receive the highest grades, participants wanted to see that student "go the extra mile"	**Perception of Student Behaviour** "Everyone bumps 1% for those strong students!"	Participants believed "teachers (and everyone else in the world) definitely make judgements based on perceptions"

Negative student behaviours are not punished per se, but positive behaviour often results in a report card mark increase

The participants were advised to get to know their students as learners in order to

differentiate instruction and be an effective educator. However, to accomplish this

task, and not get to know their students as people, was unrealistic. Consequently,

how a participant interpreted a student's attitude towards his or her own education

inevitably entered the evaluation process and a participant's professional

judgement. Although including such aspects in the process is a shadowed practice,

the participants felt it was justifiable. Providing an objective explanation for why

student behaviour should be part of the assessment process usually involved the

concept of the need to reflect on the whole student. This idea is further explored

in the next category.

3.5 Holistic Achievement

In this category, we see how the participants not only conceptualized

overall student achievement when grading, but how the respondents became part

of the evaluation construct. Participants reflected on a student's "whole"

achievement when equating this achievement to a percentage grade within the

framework of the local guidelines as well as themselves as educators. All that is

known about the evidence of learning at the time of writing the report card is

brought together. The ability to make a holistic account of a student's level of

achievement is a necessary component of an educator's professionalism. The

combination of training, experience, and the intent to act in the best interest of the

student are essential skills. Furthermore, recognizing a student's achievement

comes as a result of interaction throughout the learning process. Harry pointed out

teachers are "people who understand what their students can achieve the most,

they are the ones who are most familiar with their students; who their students are

and what they can do." Holistic Achievement conceptualized not only how

educators consider academic achievement of students, but elaborates the

importance of how personal views on evaluation were embedded in the overall process.   Although the participants expressed a wide range of views on evaluation, there was one clear pattern: the participants see their students not as things that are subject to guidelines, but people with whom the learning experience is shared. How well students did in a course could be a reflection of the effectiveness of a participant's instruction. In the meantime, participants generally interpreted being fair and equitable as being as generous as possible, even at the price of transparency. The participants did not simply make positive grade adjustments, limit punitive measures, and pass borderline students because they are told to, but because they wanted to. Holistic Assessment demonstrated an emotional need to see students do well as an indispensible part of informed professional judgement, but unfortunately is usually expressed as Heuristic Assessment.

3.5.1 "You think holistically about the child"

All participants were asked to provide a personal definition of professional judgement. Miguel identified it as "the best picture that [teachers] have of that student." In Sally's words, "You think holistically about the child. For me, that's what professional judgement is." Likewise, Lorrie stated, "when dealing with a student, you're thinking about the whole child in order to make an informed decision about something." These definitions neatly summarized the one used by the OME: it is about taking all the knowledge associated with assessment practices and reflecting holistically on the individual student. However, how did teachers conceptualize student accomplishments holistically and convert this idea into a

percentage grade? Furthermore, how does one explain the process of professional

judgement as understanding holistic achievement? Perhaps it is best to break the

single term down into two separate components: professional(ism) and (making)

judgement(s). If we look at how educators see themselves and act as

professionals, we can better understand how the judgements were justified and

what they look like in practice.

The OME's definition of professional judgement is described in two

sentences. The first sentence stated that substantial knowledge of good assessment

practices would lead to informed decisions. In practice, these informed decisions

should become further refined with continuous reflection and repetition. The

definition included a statement on the importance of awareness of "context" to

guide their judgements (Ontario, 2010a, p.152). As mentioned previously, Clarke

(2005; 2014) insisted that context was inseparable from a situation made of social

worlds and arenas. Professional judgement allowed participants to process such

conditions and interactions inherent in their local situations. As Murray pointed

out, "professional judgement is, basically, within a certain context, an ethical

context, professional context, is using your judgement, to, again, make a decision,

based on all these factors." Professional judgement as determining holistic

achievement is knowledge and observation in action in order to comprehend the

whole situation. The administration does not have a guideline for everything, so

professional judgement empowers educators to make decisions on what is

perceived to be the correct course of action, and have confidence in these choices.

We can see this notion in George's statement: "There is a recognition that… you

are somebody who has been formally trained. So in grey area, in situations where it seems what's been written in the book is not applicable to real life situations… it is up to that professional to utilize their judgement to do what they feel is best in the student's interest."

Doing what "is best in the student's interest," based on the responses from the participants, is at the forefront of determining holistic achievement. Knowing what is best was strongly connected to the assessment skills of the respondents. To Derek, professional judgement is his personal "philosophy of education" which guided him in his assessment decisions. Lisa referred to "credentials" as creating the ability to "exercise professional judgement," as Jerry spoke of his "qualifications." The Ontario curriculum is not based on closed-ended questions with right and wrong answers that simply require a referee. It is based on the desire to develop the critical thinking skills of students, and in most cases there are multiple ways to demonstrate evidence of learning. Teachers need to translate the curriculum documents to build daily lessons, figure out the best way to check for evidence of learning, design and evaluate assignments, and motivate students to be successful – all while reflecting on factors such as the Fundamental Principles and important matters such reliability and validity.

According to Clarke (2005), everything is intertwined. Those within the situation can see it, but cannot escape it; the observer and the situation coexist. In regards to the matter at hand, participants evaluated student achievement, but as the classroom teachers, they are part of the evaluation equation. To evaluate properly, participants needed to reflect on their place within the situation. For

instance, Catherine said professional judgement is "a matter of adapting yourself to the level of the course" and recognizing what students were capable of at a particular grade level. Determining holistic achievement involved more than thinking about the whole student, but included teachers reflecting about themselves. This statement is simply another way to explain the "systematic thinking process" of professional judgement that addresses "ongoing reflection and self-correction" (Ontario, 2010a, p.152). The grade may be a symbolic representation of student achievement, but it was the participants' understanding of that achievement that created the material construction. Dirk pointed out grading is "Utilizing everything that you've learned… not just pedagogically speaking, this could be life experience, basically the ability to take all of the objective facts and then apply your subjective interpretation to what you've collectively gathered… it means applying your own opinion that you can still back up with facts." Denise, as she began her teaching career, identified professional judgement as her "responsibility of figuring out the dynamic that is my classroom… to figure out the abilities of my students and use evaluation in that sense." As participants assess, they brought themselves into the construction of evaluation as they processed the grade as a product of learning. Holistic achievement is in the spirit of triangulation as described in *Growing Success*. It takes all evidence of learning to make evaluation decisions. However, there are still concerns over reliability and validity in the implementation of Holistic Achievement.

In short, the assessment process begins with the participant. Granted, the process cannot happen without the presence of student achievement, but the first step taken to understand this achievement is by the professional knowledge of the participants. We have seen there are guidelines to follow, but how a participant understands the process is where assessment commences. Furthermore, since every student is different, and no two assessments are exactly alike, the participant must to use his or her knowledge within the situation to make a judgement. Thus, as we have seen repeatedly, there is a balancing act to reach the objective of determining the final grade, where the assessor is as much as part of the process as the student. Consequently, a report card grade contains the residual beliefs of the assessor as an inseparable byproduct of informed professional judgement.

3.5.2 "Going beyond what the data [show]"

Final report card grades are an example of how a post-positivist framework can be interpretivist in nature. The participants are asked to judge overall achievement using external, provincial standards. However, tension was noted. All participants alluded to wandering from the intentions of *Growing Success* and/or local guidelines in order to act in the student's best interest, or at least to take advantage of the room they are given by the administration to manoeuver. James explained, "Teachers have to take their professionalism into account and make fair and informed decisions that might not be specifically outlined in *Growing Success* or in assessment guidelines. I mean, there's criteria, but teachers have the right to make a decision that might deviate from someone else." That is, teachers see themselves as individuals evaluating individuals. Professionalism

includes "the right" to make reasonable evaluation decisions that are "not specifically outlined." Likewise, Harry stated professional judgement "means going beyond what the data shows [sic]. And going beyond what the curriculum says, or maybe expanding on it, but also staying within the knowledge of trusted educators." Or as Oscar said, it is "the ability for teachers to provide an assessment of a student that is inclusive of, but not entirely based upon, the marked assessments during the course." It is not that the participants were purposely disregarding policy, but doing what they felt was proper within their notion of professionalism. Policy may provide direction, but it does not know, and cannot know, the individual student; the classroom teacher does. For the respondents, part of being a professional was to feel trusted in the assessment decisions that were made. On that note, Eileen described professional judgement as "judging a student fairly and have some freedom or flexibility to be able to determine not just based on numbers, but based on the whole student and their performance and what you personally know about them and their work habits to give them a mark that reflects their true performance as a student in your class." Again, we see the connection between using professional knowledge to make a value decision on the whole student. Making these decisions requires a pliable environment. We have already seen that the administration has specific guidelines and was a major influence on professional judgement. Furthermore, we have seen evidence that the participants generally support and follow both shared and shadowed guidelines. But there were always the grey areas George mentioned. It then became a matter of thinking about what is fair, equitable, and, ideally, transparent, for the individual student when determining a report card grade. Larry

noted, "at the end of the day, who knows the students best but the teachers?...professional judgement means teachers are going to know more about the capabilities of their students… it's not something that can be standardized, right?"

Larry's rhetorical question was poignant for this study: if professional judgement is an antithesis of standardization, but the provincial standards name professional judgement as its heart, then what consequences occur when it is applied in real life? When the participants said that professional judgement meant going beyond the data in order to assess the whole student, what did it look like and what does it mean for the concept of professional judgement? We already have evidence that the participants tended to follow local guidelines as they applied to the situation, which is tied in to recognizing the human and social aspects of assessment. However, we still need to establish quantitatively what professional judgement means in practice because that is where it turns into Heuristic Assessment.

3.5.3 "It ends up in the mark being bumped up more often than not"

When determining a final percentage grade, Jordan explained that applying professional judgement meant "determining trends" in the student's achievement; it was to conceptualize achievement based not just on a student's most consistent work, but the direction he or she was heading in achievement wise. The need to determine consistent achievement was a key instruction in *Growing Success*. However, the OME did not advocate one particular type of calculation.

Furthermore, nowhere in *Growing Success* is the word "adjustment" used. Hence, we can note a vital hybrid of shared and shadowed practices: "mark adjustments." It has already been explained that most of the participants were instructed to use software to determine a student's average. Also, all participants have made adjustments in some form or another, be it through administration guidelines, correcting punitive measures, and/or perception of a student's behaviour. Consequently, when it came to professional judgement as an action, the participants generally thought of the term as the method of justifying positive adjustments. As Greg claimed, "In practice [professional judgement] really means the mark going up. I really think that when people use professional judgement, it ends up in the mark being bumped up more often than not." This sentiment, at least to a degree, was shared by all participants. As previously stated, professional judgement is the qualitative element in a mostly quantitative process. Most participants used jargon for adjustments, such as "tweak", "boost," and the popular "bump." Again, it was rare to reduce a student's software-created grade. For instance, Smitty said she could "potentially… mark them down. But inevitably I mark them up." Participants were more interested in finding ways to assign the best grade possible without misrepresenting actual achievement. After all, when forced to pinpoint a percentage grade, most would rather err on the side of generosity. However, there appeared to be a general lack of awareness of how these adjustments impacted the reliability and validity of report card grades.

Jerry explained, when he is determining the report card grade of an individual student: "You know that the student… had very poor language skills,

but they were able to somehow display the learning to you, so maybe their Communication mark wasn't so good… that would be another reason that I would consider [to increase the mark]." To clarify why this evidence is in this category and not the previous one on the section regarding effort, the point here is the participant was reflecting on not just who the student was and whether or not s/he appeared to try, but what was actually accomplished consistently despite "obstacles." Furthermore, participants recognized their value judgements could have life-long implications for students. As a result, there was a tendency to find ways to justify an increase to the percentage grade.

When participants did decide to adjust the grade, what was considered normal? Helen stated she would go as "high as 5%, but not beyond." Harry would also consider a range of plus or minus 5%, but almost always favoured an increase. However, one of this study's findings was an axiology to professional judgement where *one percent does not always equal one percent*. For example, for most of the respondents, to increase a student's mark from 80% to 84% was pushing the comfort zone, even though both marks are the lower end of Level Four. At the same time, there was no issue with changing a student's mark from 46% to 50% in order to pass. Furthermore, in most cases a one percent increase was seen as innocuous, but in the case of changing a student's mark from 99% to 100% would most likely require a meeting between the participant and department head. Generally speaking, most participants indicated they would be fine with an adjustment of 1-3%. Murray points out "if a Level 3 is 74-77%, and a student consistently gets a Level 3, but the computer is reporting an overall score of 74%,

why not increase it to 77%? It is still a Level 3." In short, the higher the overall grade, the more difficult it became to justify an adjustment. At the same time, the ideas of reliability and validity did not seem to enter the process to determine an adjustment. However, the ideas of being fair and equitable certainly did. Since Heuristic Assessment does not concern itself with reliability and validity, but is interested in what is fair and equitable even at the price of transparency, this trend in adjustment practices should not be surprising.

As with all aspects of assessment, there were no black and white answers – but only a range of variance. When it came to breaking down the actual act of was the understood to be "professional judgement," it usually meant a mark increase. We have already seen participants working at schools with tight assessment regiments allowing students finishing with 40%-49% to be increased to a pass, or instructing teachers to round up grades. More liberal administrations allowed the participants to make the call, which usually results in an increase of 1-3%, especially if it does not actually change the achievement level based on the provincial chart. There was evidence of participants increasing a passing grade as much as 5%. However, the higher the grade, the less likely the student will receive an increase. When reflecting on the whole student, participants tended to find ways to be generous so they did not under evaluate achievement at the risk of over estimating.

3.5.4 "My reasons are rather personal…"

We previously discussed some aspects of professional judgement that were more procedural rather than the result of a thinking process, such as rounding up grades. Miguel's administrator suggested a few percentage marks was not statistically significant. However, to automatically round up a grade usually means the student will go up by a level of achievement. In other words, 69% is a Level Two, denoting consistently demonstrating some achievement of the overall curriculum expectations, whereas 70% is a Level Three, indicating the curriculum expectations are consistently met. Even though informed professional judgement should determine which descriptor best represented a student's true achievement, participants are more likely to mechanically increase. Oliver, while explaining why he always raised a mark ending in nine, said, "My reasons are rather personal. I finished my OAC average with a 79, and I remember feeling so upset that no teacher would bump me up just a little bit to get me to be an Ontario Scholar. So I guess it's a little personal, but I will never give a kid 79!" Note the evidence here of the emotional perspective of grading. How the student may react to the grade can enter into the decision process.

Again, holistic assessment was the aspect of professional judgement that not only called upon the participant to assess student achievement as a whole, but involved reflecting on the student as a person. This reflection brought in the emotional side of assessment, for both the student and the participant. The participant may reflect on how the student will react to the grade, and adjust accordingly. There was also the sense that the participant wanted to be

comfortable with the grade as well. Such aspects are not addressed by policy, but appeared to be a significant part of assessment in practice.

3.5.5. "You better be able to give a very good explanation as to why that jump has happened!"

Since assessment, particularly professional judgement, is subjective, attempting to explain grading to a student and/or parent/guardian can be challenging for an educator. Moreover, to admit to having some control over pinpointing a percentage grade, and attempting to explicate the thinking process behind considering student products, conversations, and observations, can sound cavalier. Consequently, there could be an invidious reaction from the other party and accusations of mendacity, so the participants often engaged in shadowed practices. The participants tended to consider all the evidence of learning and evaluate the student holistically, but it needed to be done in such a way that the grade can be backed up objectively, i.e., explained in a way that makes sense to someone who is not aware of the intricacies of assessment. Because local guidelines may also place strict limits on adjustments, sometimes participants needed to act covertly. In a previous example, we saw Winnie, Larry, and Lisa change entries in the software to efface prior poor achievement because they feel the students earned it.

Nevertheless, all the participants report a high degree of confidence in their ability to explain how they determined a grade. The explanation was based on local guidelines, but if the grade was further scrutinized by a student or parent – then the educator needed to be able to explain his professional judgement. Corey

stated, "I think it's important that you are able to explain, justify, your judgement, if you are ever questioned about why that something is showing up the way that it is... you better be able to give a very good explanation as to why that jump has happened!" Corey's thoughts were similar to several other participants' who suggested that if a mark adjustment could be justified to a colleague or department head, it was usually acceptable. However, the participants admitted they tend to avoid revealing this information to students and parents/guardians. James said, "it is fair to say that the method of calculating final grades is very rarely discussed with students and/or parents unless they expressly ask for an explanation." Again, the participants can base their explanation on policy, but once they get into detail, the process can sound too subjective. Corey's example showed if asked directly, an educator must be prepared to explain. At the same time, the respondents preferred to make their adjustments secretly, even if it presented transparency issues. As Derek said, he does not "articulate or justify an adjustment because others [students] will just assume that's the grade that was legitimately earned." Experience informs participants of what is appropriate when it comes to mark adjustments and how to best explain them.

A recurring theme to the analysis is how the respondents take great care to be fair and equitable to their students, but often at the price of transparency. At the same time, they avoid assessment decisions that are clearly unjustifiable. The main issue is in the difficulty of explaining how assessment works to an outsider. In other words, anyone, including parents and students, who does not understand Ontario policy, or has no insight into how difficult it is to grade a student on an

achievement chart, then translate the most consistent/more recent achievement to an exact percentage grade, will have a challenging time understanding the subjective nature of assessment and evaluation. Furthermore, the participants are asked by policy to note any evidence of learning that may have occurred during a conversation and/or observation and not just student products. If it is revealed that a participant does have it within his or her discretion to add a few percentage points, then this action may become expected. As a result, the participants try to avoid talking about the specifics of the assessment process.

3.5.6 "I'm constantly comparing students"

The OME's definition of professional judgement suggested it is something that evolves with experience. Although more experience should aid the decision-making process, experience alone does not constitute professional judgement. If it did, then new educators would have no professional judgement because they have no teaching experience. Again, professional judgement is not just a general philosophy of evaluation, but allows for the most appropriate holistic decision a teacher can make at a given point in time. Decisions must be made, even if an educator has limited experience to draw upon. Although Ontario is a criterion-referenced system, another legacy from normative-referencing seen in the data was comparing student work to establish a framework. Denise confessed that comparison helped her through her first year of assessing students. However, even with years of teaching experience, Larry said, "I'm constantly comparing students. Because you need a frame of reference, right? … Comparing students, for me, is a

way to get to know their abilities better, which totally helps with my assessment of learning."

Comparing students did not mean any of the participants are using normative referencing to establish quotas (i.e., a conscious equal distribution of the four different levels in a single class). However, if two students have a consistent achievement of Level Two, should they receive the same percentage grade, or should one get 64% and the other 66% by comparing their work? George admitted to comparing students, stating assessment can be "overwhelming; I don't think people know how tough marking is, especially in the Social Sciences." When conceptualizing the "whole" student, comparing students appeared to be a means to an end. Derek explained, when an educator is new, "you tend to compare students with their peers at that time. I think that that's unfair, but I think that it does happen, because that's the most accurate representation of what's taking place." Derek went to say that "over the years, cumulative knowledge" established a more proficient capability to determine grades. He was not just comparing two students in the same class, but all the students he has taught. It may be another shadowed practice and an example of Heuristic Assessment, but to the participants, it was a pragmatic approach to pinpointing a percentage grade.

In addition to considering effort, comparing students could also help with what Catherine called the "ambiguity" of Level 4. When asked about the difference between an 85% and a 95%, Harry stated, "I think there's a distinct difference between both of those grades. Because a 95 is nearly perfect, nearly reaching all of the expectations and outcomes with very few mistakes. An 85 is

above average, but not quite reaching the level of perfection the 95 did." In other words, the progressive tiers of Level 4 cannot be explained without referring to exemplars that demonstrate the differences. To base assessment on the "whole" student, it helped to reflect on other students and their unique characteristics to appreciate the concept of achievement. However, there was the sense that gauging a student's overall achievement in this manner was also an example of Heuristic Assessment.

Professional judgement should incorporate all evidence and procedures to determine the most accurate grade possible. However, what does not seem to be part of the conversation was recognizing a margin of error. Participants are expected to precisely rate a student's most consistent level of achievement on a scale of 0-100. Instead of thinking in terms of margin or error, they commonly err on the side of caution by increasing a computer-generated grade. Policy is vague on such aspects of evaluation, so administrations and respondents fill in the blanks. As a participant builds his or her professional judgement, grading confidence increases, but there are still intangibles such as pinpointing student achievement. Consequently, participants reflected on the whole student, including comparisons to other students, and act within the local guidelines to determine overall achievement as part of Heuristic Assessment.

3.5.7 Holistic Achievement and Professional Judgement Summary

Chart 3.4: Professional Judgement and Holistic Achievement shows how different factors are used to determine a percentage grade. Due to the difficulty of

this balancing act, particularly because of the emotions involved, it makes sense

that the participants were as generous as possible. When addressed directly,

accuracy on a 100-point scale is incredibly challenging, especially when this

decision has life-long consequences. Therefore, it is understandable why a

participant would want to give a mark a slight, positive adjustment, or find ways to

Chart 3.4: Professional Judgement and Holistic Achievement

Participants indicated "You think holistically about the child. For me, that's what professional judgement is"

Determining overall achievement involved "Going beyond what the data shows"

Professional judgement tended to mean the "mark being bumped up more often than not"

Holistic Achievement

"[participants were] people who understand what their students can achieve most"

Evaluation can become "rather personal"

Participants needed to be able to "give a very good explanation" for the adjustment

The participants not only use professional judgement to evaluate achievement, but the teacher becomes part of the construct. Grades are a reflection of the learning process itself

Although participants knew they should use criterion-referencing, they are "constantly comparing students"

pass a student. However, with so many shared and shadowed practices, how can a

participant be confident in a decision? It has been stated the participants are

prepared to explain their decisions, even if it is a conversation that they would like

to avoid. In short, the participants gathered what they knew about assessment,

combined with a set of personal rules, such as how to compare students, to guide them through the process that may involve adjustments.

3.6 The Core Category: Heuristic Assessment

Is professional judgement merely, as Lucy says, just a "loophole" to make policy more manageable? That is, is professional judgment the means to cope with the confusion in the assessment system? We have seen four categories that shed light on what professional judgement means in practice, but the research inquiry still needs to be answered. Charmaz instructed, "Rather than aiming for theoretical generalizations, constructivist grounded theory aims for interpretive understanding" (2011, p.366). Professional judgement is a vast concept. Even to narrow it down to its role in determining report card grades still produced a situation with a wide range of variance. Still, abduction allowed for a plausible explanation for what is happening. Connections between the four categories have been discussed in passing. The people involved in the situation, namely the participants and their administrations, addressed various conditions and followed shared and shadowed practices in the attempt to influence the situation to allow for maximized student success. At the same time, the instability of the situation, caused by the competing factors such as turning policy into practice, dealing with grey areas, and the emotions surrounding high-stakes decisions, demanded that the participants create a sense of stability to accomplish the task of quantifying success as a percentage grade. How the participants create this stability is discussed here in the core category of Heuristic Assessment. As the label suggests, the core category not only explains how all the categories are linked, but

how the participants established "rules of thumb" to navigate the assessment

process. It is an interpretation of not only what the categories have in common,

but a general explanation of the process behind how the participants were able to

determine report card grades using their professional judgement. The following

subheadings further explain the bridges between the categories and the core

category to show how Heuristic Assessment, as Dirk explained, "brings everything

together." The procedures can be convoluted, but it gets the job done.

3.6.1 Negotiating Rules

Similar to how Strauss (1993) explained social worlds and negotiation, this

chapter demonstrated that determining a student's final report card grade involves

negotiating many situational rules, including the meaning of achievement.

Consequently, explaining how a student's percentage grade was determined could

easily become an exercise of byzantine reasoning. *Growing Success* provided

broad guidelines, but not step-by-step instructions. The instructions are suggested

by local administrations that interpret policy. Significant differences in

interpretation have been noted. For instance, Lucy's school had abandoned

computers and instructed teachers to use mode and professional judgement to

determine a grade. On the other hand, Lorrie's school disregarded the whole

notion of most consistent and more recent achievement and follows a pre-New

Curriculum approach to assessment. The participants tended to adapt their local

situation and usually follow the general guidelines. However, even when a school

exercised a high degree of control over teacher autonomy, as George pointed out,

there are still variables for educators to address. A department head cannot be

consulted on every decision. Therefore, the participants needed to consider the

individual student and his/her individual circumstances through the lenses of

achievement charts, curriculum expectations, fundamental principles, learning

categories, and summative assessment, with the goal of pinpointing a percentage

grade that was reliable and valid and has no margin of error. When we take a

moment to realize the impossibility of this task, we can focus on evaluation's

inherently subjective nature.

Generally speaking, the participants took individual assessments and

entered the levels into computer software to get a calculated grade. When it came

time to produce a report card, participants consulted their local guidelines on what

to do next. At the same time, participants used Heuristic Assessment to decide on

adjustments. Greg spoke of professional judgement as a euphemism for the

pressure an administration could place on teachers to maximize grades, but he still

managed to have his own adjustment rules. For others like James who were given

a high degree of autonomy, he reviewed individual assessments for outliers, as

well as reflecting on evidence obtained through conversations and observations, to

inform his professional judgement and triangulate the most consistent level of

achievement. However, even with significant freedom to determine a grade, he

still needed personal rules to decide on how best to translate a level into a

percentage grade. For those like Oliver who needed to stay close to the computer-

generated average, it might just be a matter of rounding up a grade or awarding a

few extra percentage marks based on recent achievement. Even Lorrie could

usually find an extra percentage point for those she felt who earned it. Regardless

of the local situation, all the participants were reflecting on possible ways to justifiably increase a student's grade along with the local rules. This action not only benefited the student, but also appeared to have an emotional connection to the participant as a desire to see students to well.

Participants like Corey talked about how difficult it is to ignore student behaviour and its link to percentage grades. Derek made this link explicit to students. The participants did not give the impression that students receive an overall mark deduction as a form of punishment for poor performance. With the push from the local administration to ensure students pass the course, "misbehaving students" are likely to receive the computer-generated grade with no adjustments. However, the students who were deemed to be hard working, i.e., who noticeably display effort, were likely to receive a mark increase from participants such as Larry, Lisa and Winnie by retroactively adjusting individual marks.

However, there was general confusion regarding the rules governing punitive measures. Although policy tried to downplay punitive measures as a last resort, they have quickly re-established themselves at all the schools represented by the participants. Most of the participants such as Helen portray punitive measures as a veiled threat to students in order to gather evidence of learning, justified along the lines that it is the way "the real world" works. However, at schools at which Catherine and Jordan work, there are harsh penalties for missing assignments. Recorded punitive measures – and the need to have a correlation between the computer and report card grade – put these participants in an awkward

position as any zeros must be included in the average. Again, a participant's emotions lead them to not want to punish the student, but sometimes the administration makes it unavoidable. In such cases, the participant can either adapt or leave the school like Eileen did.

Nevertheless, guidelines were meant to help more than to hinder, keeping the curriculum expectations in mind, advising participants how to look beyond behaviour, or to better understand what to do when there is missing evidence of learning in the equation. As Miguel said, ultimately educators need to focus on the best of what a student has shown. The best is still a relative term, and needs rules to help explain it. It short, the general rule lead participants to reflect on the evidence that had been gathered throughout a course and is recorded as holistic achievement. Even in schools where it is more difficult to include conversations and observations in the grade, when a participant knows a student's strengths and weaknesses, the student can be better coached for products that will be added to the portfolio of work.

3.6.2 Need for Stability

In the chaotic social world of school, the participants needed to make sense of the situation and create stability (Clarke, 2005). Explaining the journey of how curriculum expectations eventually become a grade on the report card quickly obfuscated the evaluation process, even for trained educators. By making sense of the process, a participant could be a more confident assessor, especially in cases such as Smitty who had decades of experience. When participants feel they

understand the procedure, they achieve a sense of stability. Sally explained professional judgement was about reflecting on the whole student. Should a participant experience dissonance with the process, Heuristic Assessment could provide a reminder to retreat to stable ground by finding ways to look at the big picture. However, what does a participant do when the big picture was still confusing? Although the grade was based on achievement of the curriculum expectations and the local rules, these expectations in turn must align with the fundamental principles of being fair, transparent, and equitable.

Heuristic Assessment allowed the participants to bring together a way to evaluate curriculum expectations and the fundamental principles. For example, the idea of being fair was central to Eileen's understanding of professional judgement. We see the aspiration to be fair in all the supporting categories: by following guidelines that were applicable to all students, by not dismissing a student because of certain behaviour, to provide multiple opportunities to make up for missing work, and to look at a student holistically. Also, being fair was to be mindful of the group. When evaluating, Larry actively compared students not in the sense of normative referencing, but to feel confident that no one student is treated better than another.

Many of the participants used the term equitable interchangeably with fairness. For example, when a student was evaluated holistically, being fair and equitable are close to one another. In practice, however, participants applied the difference. In short, being fair removed bias from the evaluation system and placed a check on favouritism and/or discrimination. However, it did not mean

treating all students the same. As Harry explained, it is the mission of the

classroom teacher to get to know students and how they learn. Some students

have special needs, so guidelines allow for accommodations. Outside factors

beyond a student's control could be the source of negative behaviour, including

missing assignments, placing the student at risk of performing poorly. Being

equitable took these outside factors into consideration.

However, all the categories, including Heuristic Assessment, can be

challenged by the principle of transparency. As long as it is explained how

students are assessed, e.g., with a rubric, the bare minimum of transparency is

achieved. Transparency is strengthened when a participant uses both the

curriculum expectations and the fundamental principles to explain evaluation

procedures. However, each category has shadowed practices, usually involving

some modification to the percentage grade. This conflict did not help an already

confusing situation. Nevertheless, participants like Jerry take solace in the idea

that, while certain practices could be more transparent, they were carried out in the

student's best interest. In the feedback he provided, Oscar stressed the importance

of going beyond curriculum documents and the genuine desire to care for the

success of his students. Nevertheless, the Ontario system can do better, in the

spirit of the fundamental principles than relying on Heuristic Assessment.

3.6.3 Providing the "Right" Assessment

Heuristic Assessment is a modus operandi for the participants. By this

point, it should be clear that participants were taking the guidelines seriously, and

balancing both the curriculum expectations and the fundamental principles in their evaluations. They tried to limit favouritism to those who were hard working, and not to dismiss those who normally appear uninterested but delivered quality student products. Generally speaking, the reality of the situation is less a matter of extremes. High-achieving students tend to remain high achieving, so it is the weaker students who require more of the participant's attention. Some of the participants spoke of the satisfaction of reaching through to students and making them realize their potential, but more commonly, it is a matter of getting the struggling students through the course. Nevertheless, the feeling that every student earned a passing grade is in itself a reward. Some participants like Lisa speak of class average targets, but it was more of a relief to not to feel that a student was given a gift pass, and is more likely to be prepared for the future. Ideally, participants are left with the feeling that the "right" assessment was provided. Pringle (2007, p.147) believed "Persons are actively seeking to make sense of experiences as they strive to meet their needs and to solve problems that confront them." In order to understand the situation, they need to make connections between the unknown and the known in an ever-changing social world. The participants in this study were also interested in making sense of their social world. This world was already constructed when they started their profession, and came to understand its rules through interaction. The better they understand it, the more they can learn. However, the infrastructure of school also interacts with the participants, and can change as well. By accepting new ideas about assessment and evaluation, further improvements can be made.

3.6.7 Heuristic Assessment and Professional Judgement Summary

Chart 3.5: Heuristic Assessment and Professional Judgement illustrates

only benefit student grades, but also help the participants cope with a confusing

system. The procedures are far from flawless, but they reflect the reality of the

Chart 3.5: Heuristic Assessment and Professional Judgement
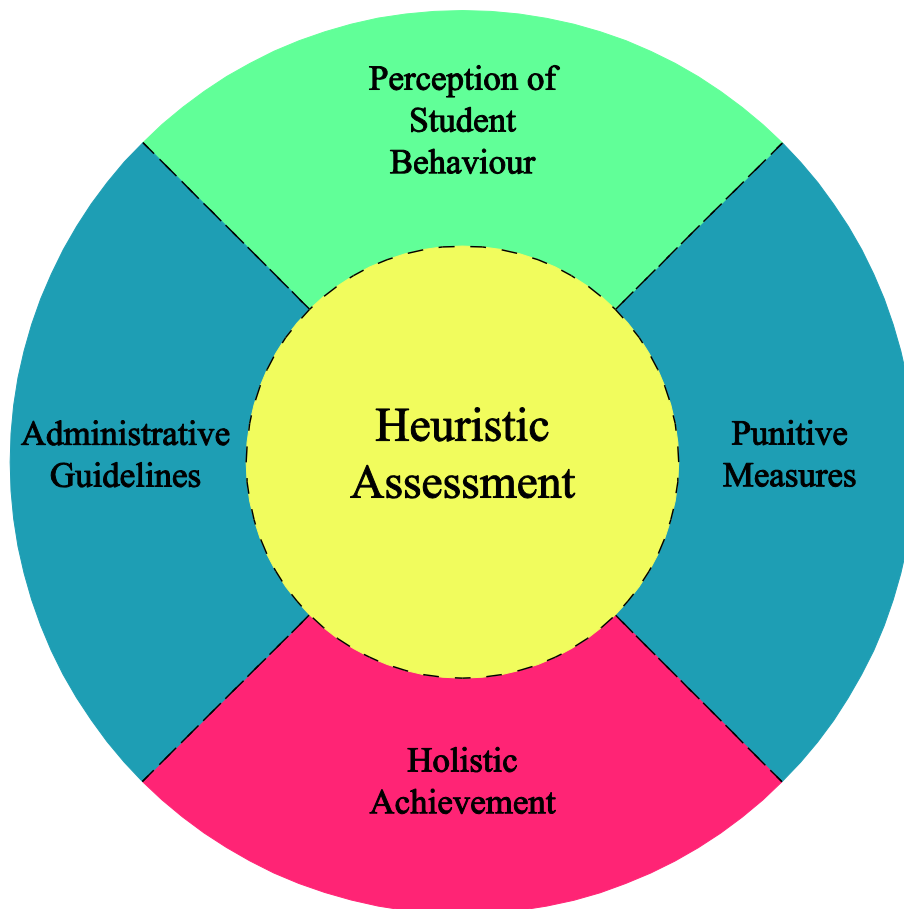
how the categories of Administrative Guidelines, surround the core category of

Heuristic Assessment. The chart contains porous boundaries represent how

different elements of professional judgement manifest in the evaluation process.

Chart 1.2 in the Introduction (see p.45) showed different external and internal

factors professional judgement needs to consider according to policy. Chart 3.5 could be placed at the centre of that chart to demonstrate how the factors are processed and conceptualized in practice. The situation is always one of disorder, but social processes find a way to establish a methodology for the actors to understand, in their own way, how the pieces fit by negotiating rules to establish stability and to provide the "right" evaluation. In regards to professional judgement in practice, this methodology is Heuristic Assessment. It is the way, as Murray put it, to reduce assessment into a matter of rules of context, from the point of view of the assessor. Heuristic Assessment makes sense of the overall process by bringing together emotion and reason, policy and practice. However, it also mixes shared and shadowed practices, interferes with the fundamental principles in the interest of upholding them, and brings reliability and validity issues into the Ontario evaluation process.

3.7 Constructing Heuristic Assessment Summary

This study found that Ontario teacher participants applied their professional judgement to maximize student success by applying not only local guidelines, but also ones that are personally created in the form of Heuristic Assessment. These personal rules based both personal reasoning that blends emotion and logic, not school and the inescapable shortcomings of the concept of evaluation. Addressing these shortcomings could correct the problems associated with Heuristic Assessment, and make the system of evaluation more reliable and valid by improving the practice of informed professional judgement.

Determining a report card grade should be the logical conclusion of assessing a student's achievement using reliable and valid means. However, there appears to be interference from multiple sources: the role of non-achievement factors such as punitive measures and student behaviour, the way assessment and evaluation policy is explained and upheld at the local level, and the human factors involved of evaluating someone with whom a learning experience has been shared – especially when this judgement of achievement has real-world consequences. A review of the literature, discussed in the next chapter, will show how the challenges faced by the participants in this study are found in many other systems of assessment and evaluation. By taking into account these common obstacles, recommendations can be made for further effective reform.

CHAPTER FOUR: LITERATURE REVIEW AND DISCUSSION

4.1 Introduction

The preceding chapter demonstrated how the participants applied their understanding of professional judgement to final report card grades in Ontario Secondary Schools. As a result, answers to the research inquiries concerning the application of professional judgement were constructed. Respondents engaged in what was termed Heuristic Assessment: when deciding upon a student's overall percentage grade, they applied personalized procedures. These procedures were based on interpretation of provincial and local policy, on learning and curriculum expectations, and on achievement and non-achievement factors. General trends included erring on the side of caution, such as rounding-up grades, adding percentage marks for good behaviour, and passing borderline students. In short, participants attempted to simplify evaluation in a way that not only made sense to them – but was based on a perception of what was fair and equitable even if at the price of transparency. Such findings were grounded in the voices of the respondents, however, this explanation is only one possible view. Therefore, this interpretation must be discussed with regards to the extant literature for validation purposes. Black and Wiliam (2003) show how educational research can be shambolic. A review of the literature establishes where this study fits in the field of knowledge (Schwandt, 1996; Miller and Fredericks, 1999).

Although there are a few recent sources on evaluation practices in Ontario, research into professional judgement as a whole is limited. The review included studies from Australia, Europe, and North America in order to generate a detailed

discussion. Furthermore, authors tend to discuss professional judgement in connection with a larger range of topics, including issues of reliability and validity, non-achievement factors and evaluation, and formative and summative assessment. Since this study also addressed these points and their connection to report card grading, a thorough review has been provided.

Most of the sources used similar methods: a survey/questionnaire of educators, with the possibility of follow-up interviews (e.g., Earl, et al., 2010). Only a few studies mentioned constant comparison (e.g., Hay and MacDonald, 2008). Discussion of methodology was almost non-existent, with many assessment authors alluding to their own research and literature review instead of discussing epistemology (e.g., Gardner, et al., 2010). As a result, the discussion in this chapter will take a thematic approach to explore the various topics raised in the research.

A careful examination of the sources used in *Growing Success* and complimentary literature showed that the actions of the participants were essentially consistent with comparable practice in Ontario and other education systems. It can be noted that the New Curriculum has not solved some of the fundamental challenges to educational evaluation, such as issues concerning the reliability and validity of grades and the difference between process and product. By examining these issues, recommendations can be made for further improvements to provincial policy to resolve certain assessment dilemmas, thus making a contribution to the field of knowledge.

4.2 The Challenges of Researching Professional Judgement

Countless studies are available on student evaluation. Gardner (2012a, p.2) observed, "The documentation on the various assessment policies, practices and theories could conceivably fill whole libraries." As discussed in the Methodology and Methods chapter, a limited review of subject literature was completed as part of the study's proposal. A more comprehensive review was deferred until the completion of the CGT analysis so the findings would not be influenced by existing ideas. However, the detailed review was challenging; database searches for keywords such as "professional/teacher" and "judgement/judgment" produced relatively few results. Searches for "Ontario report card" provided thousands of results ranging from environmental issues to medical, but almost nothing on education.

In personal correspondences, assessment authors and consultants Cooper (2010) and O'Conner (2010) stated that the term professional judgement was commonly used but not usually explained. In a study released just prior to *Growing* Success, Earl, et al. (2010) surveyed 5905 Ontario educators from all 72 boards, with follow-up interviews with 72 teachers and nine principals from nine schools across the province. The researchers noted that the participants were aware that professional judgement was needed as part of evaluation, but did not clearly understand what the term meant. Frary, Cross and Weber (1993), in their questionnaire-based study on how 536 Virginian Secondary School teachers viewed assessment practices, also noted a lack of awareness of professional judgement. Other researchers have commented on the lack of quality research on

professional, or teacher, judgement (Stiggins and Bridgeford, 1985; Stiggins, Frisbie and Griswold, 1989). Reiman and Johnson (2003, p.4) conducted a meta-analysis of literature on assessment practices of American professors and noted that although the idea of teacher judgement dated back to John Dewey (1904), the term was "poorly defined in the literature" and there was "little research" on its proper use. Allal (1988) interviewed 45 Geneva Elementary School educators from three schools and reached a similar conclusion. Allal (2013) found greater awareness of professional judgement in a follow-up study of ten Grade 6 teachers from ten Swiss schools, but there was still more investigation needed. Although more attention has recently been paid to the use of formative assessment and professional judgement, how to properly assess a report card grade is often overlooked (Smith, 2003; Freebody and Wyatt-Smith, 2004; Wyatt-Smith, Klenowski and Gunn, 2010; Wyatt-Smith and Klenowski, 2013).

In order to find discussions on professional judgement, a wide range of assessment literature needed to be consulted. When sources were identified, they were compared to the research inquiries, i.e., what does it mean in regards to final evaluations, how does it work in practice, and how can this understanding help to improve its application? Sources tended to discuss professional judgement not in terms of how educators defined it, but how it guided actions and the consequences that followed. Teacher judgement is needed for an evaluation to have authenticity (Newmann, 1990). For example, Hoge and Coladarci (1989) performed a literature review on teacher judgement, but focused more on aspects of validity and reliability. Gardner, et al. (2010) also reviewed assessment literature, and

suggested evaluation always requires some level of professional judgement. Wyatt-Smith and Klenowski (2013), after multiple interviews with 89 Elementary and Intermediate educators from 49 Australian schools, also said that informed professional judgement will always be needed regardless of criteria as assessment is a fluid process. This view was also expressed in Klenowski and Wyatt-Smith's 2013 textbook on assessment. As part of his research into teaching practices, Polkinghorne (2004) spoke of teacher instinct called phronesis, or embodied reasoning, which comes from experience and disciplined intuition, or prudence. In Cooper's observations on assessment (2007; 2011), he pointed out that professional judgement is the act of grading, based on knowledge and experience, to make tenable decisions. This message is similar to what is found in Airasian (1994) in his textbook on classroom assessment. Assessment should not be rigid and teachers should augment the criteria as they see fit (Hammond, 1996; Cooksey, Freebody and Wyatt-Smith, 2007). The participants demonstrated that there are grey areas in the situation; the real world is complex and there will always be trials in assessment practices (Sadler, 1985; 2009).

When discussing assessment, the issue of reliability and validity is often raised. More specifically, if professional judgement exposes the inherent subjective side of assessment, how can evaluation be trusted? Many authors have noted non-achievement factors tend to enter the evaluation process, and that this interferes with judging the accuracy of grades and interpreting their true meaning. Allal (1988) found the procedures used by educators were a curious mix of routines and rules to collect evidence from a multiple of sources. Such decision

making is essential to teaching (Good and Brophy, 2000; Wilen, et al., 2004; Thornberg, 2007; 2009; McMillan, 2008; Popham, 2009). However, educators can have difficulty matching achievement to academic standards (Wyatt-Smith and Klenowski, 2013). Zhu and Urhahne (2015) spoke with sixteen Chinese English teachers, representing 555 students, and determined that the educators tended to increase grades because cultural pressures demanded higher grades. In a recent literature review, Goldstein (2015) found that teacher value judgements tended to hurt the validity of assessments. These findings should not suggest that educators are apathetic about evaluation accuracy but reflects the difficulty of maintaining validity and reliability in the face of so many other pressures within the assessment situation. As Stiggins (2005, p.283) wrote, assessment was challenging and it is "better to keep it simple, even if the results might be imperfect." This statement is an excellent summary of Heuristic Assessment.

The literature showed that the participants in this study encountered the same issues that have been written about for decades. In short, when the respondents applied Heuristic Assessment, non-achievement factors were seen as relevant, hence hurting the reliability and validity of evaluations. The actions of the respondents were understandable in regards to fairness and equity, but the lack of transparency in both shared and shadowed practices interfere with the meaning grades are meant to convey. By further examining the literature, we can better understand why Heuristic Assessment occurs, and what can be done to address this problematic situation and improve informed professional judgement.

4.3 Literature Overview of *Growing Success*

Section 1.6 only mentioned *Growing Success*'s sources in passing in order
to focus on the formation of the research project. To understand how *Growing
Success* envisioned professional judgement, it is appropriate to review the sources
that informed its position. Of the 31 sources cited by *Growing Success*, sixteen are
past OME documents. Again, the policy is more of a clarification of assessment
guidelines and did not dramatically depart from its progenitors. However, these
previous documents did not discuss professional judgement. Only seven sources
in *Growing Success* are actually germane to this study: Sutton (1991) addressed
the challenges associated with good assessment practice based on the British
classroom experience. Similarly, Harlen and Deakin Crick (2003) and Harlen
(2006) both conducted literature reviews on summative assessment. A Joint
Advisory Committee (1993) was formed of representatives from various Canadian
educational associations and teacher organizations to suggest standards for
assessment and evaluation practices in the country. Likewise, the Western and
Northern Canadian Protocol (Western, 2006), headed by educational
representatives from Canada's western provinces and northern territories, provided
suggestions for assessment as, for, and of learning. Volante (2006) and Davies
(2007) also addressed Canadian classroom assessment based on their own
research.

Since *Growing Success* lacked in-text citations, it was necessary to closely
read both the policy and its sources to identify connections. *Growing Success*
defined professional judgement as a teacher's informed decision based on

achievement evidence, assessment guidelines, and methodical analysis. This

position appears to be partially based on the Joint Advisory Committee (1993)

statement that professional judgement was essential to being fair and equitable,

which also demonstrates the strong connection between professional judgement

and the Fundamental Principles. Likewise, Western (2006) stressed the need for

professional judgement to be informed by achievement data, and to use this

information to make key assessment decisions. Similar to those who wrote about

professional judgement being a response to an assessment situation, Sutton (1991)

and Davies (2007) explained professional judgement as an action within the

assessment process. They suggested professional judgement questions what the

student is able to show, signs of where a student is struggling, and overall student

development. In other words, professional judgement plays an active role in how

evidence of learning should be elided to form an overall grade.

4.4 Professional Judgement as a Reflection of Teacher Values

In short, this study claims that informed professional judgement should be

active and found throughout the assessment process. In contrast, Heuristic

Assessment, in regards to determining final report card grades, is based more on

assumptions, emotions, and set procedures to establish a product (i.e., a grade).

The analysis demonstrated that Heuristic Assessment occurs when the participants

applied their own interpretation of evaluation policy, and included information

from non-achievement factors. Both Heuristic Assessment and informed

professional judgement belong to the same range of variance. The crucial

difference is professional judgement should be dynamic and informed from

achievement data when it comes to making assessment decisions. On the other hand, Heuristic Assessment is a static simplification of the overall evaluation process expressed as a product. This observation is not to take away from the good work of the participants, but demonstrates the human elements involved in evaluation and the personal need to make sense of the assessment situation. It is challenging to consider achievement data only, especially when reflecting on non-achievement data feels like the fair and equitable action to take even if it is not transparent. Hence, to better understand why Heuristic Assessment happens, it is important to note the personal values that educators bring into the evaluation process.

The concept of Heuristic Assessment suggests that evaluation has as much to do with the personal values of an educator as it does with Ministry guidelines. The literature has also noted a connection between educator values and assessment (McMillan, Hells ten and Klinger, 2011). This sentiment is found in Sadler (1985) in his study on the origin of evaluation criteria, O'Connor's writings on how to improve grading practices (2000; 2009), an action research and case study on Masters students and teacher moderation by Elwood and Klenowski (2002), the findings of Cummings, et al. (2006) on teacher judgement, and the previously mentioned study by Wyatt-Smith and Klenowski's (2013). In the analysis, we saw how, when participants were deciding upon a grade, they considered an edifice of what they saw as personally important. These values were usually expressed as reflections on the real world, demonstrating the contextual side of professional

judgement in the decision-making process thus shifting it towards Heuristic Assessment.

When the participants reflected on their own value system, it usually modified, or even replaced, official guidelines. This phenomenon was noted in articles on assessment discrepancies such as the Stiggins, Frisbie and Griswold (1989) influential case study of fifteen veteran American High School teachers. Likewise, Brookhart (1993) studied the responses of 84 experienced American educators in a Masters of Education program and found they generally reflected on their own value system – even though half the participants had taken a course on proper educational measurement techniques. In a literature review, Stiggins (2001b) found little was being done to prevent teacher values from finding their way into assessment practices. Stiggins (2001b) found that educators felt the actions were justified if done in the best interest of students. Other researchers have also found that teachers tend to be altruistic (Brookhart and Freeman, 1992); they are inclined to be sensitive to student anxiety towards evaluation, although positive modifications to grades tend to happen without the student's knowledge (Newton, 2005; 2007; Harlen, 2012b). An educator may have good reason to make such modifications based on professional judgement, but when such actions take place in the shadows they become a form of Heuristic Assessment. Likewise, these actions are often inline with the desires of the local administration.

4.5 Administration Oversight of Teacher Professional Judgement

Personal values aside, participants generally agreed with the administrative "limits" placed on them. After all, part of professional judgement is to recognize the correct course of action as instructed by superiors. How administrations instruct teachers on assessment has been discussed in the literature. For instance, Truog and Friedman (1996) examined the written grading policies of 53 high school teachers, followed by focused group of eight teachers. They found administrative demands did constrain teacher professional judgement and partially controlled student grades. These findings were comparable to Harris and Brown (2009), who applied a phenomenological analysis to 161 questionnaires and 26 follow-up interviews with educators from New Zealand. Cross and Frary (1999) surveyed 397 Virginian middle and high school teachers and 8664 students and found significant differences in grading practices from school to school even within the same district. How the administration explained and controlled guidelines was seen as a reason for these differences. McMillan's own views (2003) also described how internal beliefs of the educator, and the external pressures of a hegemonic administration and other stakeholders created tension in assessment practices, which corresponded to what Stiggins, Frisbie and Griswold (1989) found. Simon, et al. (2010), a case study of an Ontario French Grade 10 Math teacher, as well as Earl, et al. (2010), noted that Ontario teachers mostly follow the local administration's lead on assessment practices including those that seem to go against OME standards.

4.6 The Subjective Nature of Assessment

Assessment literature usually argues the maxim that assessment and evaluation are naturally subjective. Therefore, it is understandable that factors such as personal values and administrative control will influence how an evaluation is conducted, even when such decisions interfere with objectivity. However, with the exception of the quotation from Sutton (1991, p.2.), which described assessment as subjective and "an exercise of human communication" (cited in Ontario, 2010a, p.29), *Growing Success* appeared to downplay the subjective nature of assessment and evaluation. As we read in the first chapter of this study, there are specific guidelines for Ontario educators to follow with regards to the fundamental principles, content standards, and distinguishing between formative and summative assessment. At the same time, interpretation of much of the policy was left to the reader. The analysis demonstrated that understanding and application of policy depended on the individual school and participant. Consequently, application of the policy was incongruous. Furthermore, the combination of shared and shadowed practices interfered with the policy's intention to be transparent. These factors, summarized as Heuristic Assessment, may be the natural consequence of not confronting the subjective nature of assessment.

Harlen (1994) and Harlen and Deakin Crick (2003) used previous studies to point out that evaluation is as approximation and cannot be treated as precise. Likewise, Mislevy (1993) referred to overall grades as caricatures. In a review of the practices of Scottish educators, Hayward (2015) concluded that professional

judgement was widely misunderstood and in need of greater transparency. McMahon and Jones (2015) found a similar case with Irish Chemistry teachers. Davies (2007, p.93) included a quotation from Elbow (1986), stating, "When we give grades or comments that try for objectivity or impersonality or general validity, we are very likely – not to put too fine a point on it – to be telling lies." Davies encouraged educators to approach assessment organically and not mechanically; they should adapt to the needs of the student in order to support learning. What Davies suggested reflects the realities discussed in the analysis. The participants were trying to account for variables ranging from administrative requests to individual learning styles, to equate overall achievement on a 100-point scale as part of their informed professional judgement. As a result, they need to establish personal guidelines to accomplish this goal, albeit with antithetical values that result in Heuristic Assessment.

At the same time, teachers prefer to view their assessments as objective, even if this belief is not reflected in their practices (Frary, Cross and Weber, 1993; Shepard, 1995; McMillan, Myran and Workman, 2002; McMillan, 2001; 2003; O'Connor and Wormeli, 2011). There are many other examples in the literature of the lack of overall objectivity in assessment and evaluation. For instance, Airasian, Engemann and Gallagher (2007) explained in their grading practices textbook that grades are always incomplete. McMillan, Hellsten, and Klinger (2011) and a Black and Wiliam (2012) paper stated that different educators, even when using common standards, could reach different conclusions on the value of a student product. Kohn (1993), in his book on student punishments and rewards,

argued that grades could only pretend to be objective. Numerous other studies

have also warned that grades are prone to measurement errors (Walvoord and

Anderson 1998; Marzano, 2000; Newton, 2005; Duncan and Noonan, 2007). The

lack of objectivity makes the assessment process, including professional

judgement, mysterious (Speck, 1998). Indeed, the voices of the participants

showed that there are rarely objective directions to follow as every student's

situation is unique.

The fact that assessment is naturally subjective does not mean that

professional judgement is not dependable. For example, Hoge and Coladarci

(1989) found teacher judgement to be consistent with externally-audited tests.

Likewise, Smith's (2003) literature view and discussion that showed the academic

averages of American High School students were similar to the scores on college

entrance exams indicating accuracy in teacher judgement. A classroom teacher is

in an excellent position to understand a student's most consistent level of

achievement, something a one-time large-scale assessment cannot capture, and

conceptualize this achievement as a grade (Black, 1993; Stiggins, 1999b).

Nevertheless, the subjective nature of assessment means that educators need to

tune their professional judgement to ensure their evaluations are trustworthy

(Stiggins, 1992; Shepard, 2000a; O'Connor, 2007; McMillan, 2008; Wyatt-Smith,

Klenowski and Gunn, 2010; Wyatt-Smith and Klenowski, 2013).

4.7 The Reliability and Validity of Professional Judgement

In her paper on proper communication of student achievement, Brookhart (1999) cited the Joint Advisory Committee (1993) as a good example of why there must be reliability and validity in assessment. However, if assessment is ultimately subjective, then how can it be reliable and/or valid? In Section 1.6.5.2, it was shown that *Growing Success* emphasized the importance of assessment consistency so grades can be trusted as a basis for decisions, for example, post-secondary education, (p.2). As discussed in the first chapter of this study, *Growing Success* contained a limited discussion of reliability and validity. Policy did make it clear that evaluation should reflect the content standards, and should be designed to produce consistent information on which important decisions can be based. However, no practical examples were included. Linn (1994) observed that educators often talk about validity, but then do not follow through with action because it is difficult to demonstrate evaluation in practice. When assessment is subjective to begin with, and even its so-called objective side has its own issues, there is little wonder why educators attempt to simplify the process through practices like Heuristic Assessment.

It does not help that there is an apparent conflict between reliability and validity in regards to student assessment (Kane, 1982; Moss, 1992; Wiggins, 1993; Brookhart, 1999; Whittington, 1999; Harlen, 2005b; Newton, 2005). Obtaining adequate evidence of each curriculum expectation is time consuming. Therefore, an overall grade that is truly both reliable and valid could only be the result of painstaking evaluation for both the student and the teacher. Black and Wiliam

(2012) encouraged viewing reliability as an aspect of validity, but admit an aporia between the two. Likewise, Shepard (2000a; 2000b) believed that if an educator collects a substantial amount of evidence, the validity of the evaluation outweighs concerns about reliability.

Moss (2003) confronted validity theory and the psychometrics of reliability in an article about the qualitative research course she taught to graduate students, where she demonstrated conflicts within the traditional understanding of classroom assessment. She asked questions such as, how could different pieces of valid evidence be brought together as one piece, while still maintaining validity, if traditional psychometrics did not provide the tools for doing so? She made interesting points that are applicable to our discussion. Rich, detailed evidence helps the researcher made proper interpretations, just as the classroom teacher needs to reflect on multiple forms of achievement. She suggested a more holistic, hermeneutics examination of student achievement, including considering the overall situation. The collection of scores can provide evidence to make grading decisions, but it required professional judgement. This is not an excuse for the teacher to do whatever he or she wants, but the nature of professional judgement demands room to maneuver. Unfortunately, sometimes the result of this maneuvering is Heuristic Assessment.

In order to make summative assessment more reliable and valid, educational systems such as those found in the United States and the United Kingdom have long used standardized testing. Ontario has resisted this approach, and limits large-scale assessment. The only externally-audited Secondary School

graduation requirement is the Grade 10 Literacy Test, which students have

multiple opportunities to pass. Despite public approval in Ontario for the test

(Artuso, 2014), the system relies on the classroom teacher to determine course

grades. Overall, assessment experts support the ability of the classroom teacher to

accurately arbitrate achievement. The discussion can be noted back to Scates

(1943), who wrote an article questioning the wisdom of standardized testing based

on quantitative and qualitative criteria. Measuring achievement should not be

viewed through a scientific, positivist lens. This notion has been echoed by

Stiggins and Bridgeford (1985), Allal (1988; 2013), Hoge and Coladarci (1989),

Brookhart (2003), McMillan (2003) as well as a recent literature review by

Elwood and Murphy (2015). Researchers have noted how educators mix

interpretive and objective factors with classroom psychometrics. Therefore, there

is a distinct difference between traditional ideas of reliability and validity and

measuring student achievement. Brookhart (2003) proposed the term

"classroometrics" to differentiate classroom assessment from traditional

psychometric terms, but it does not appear to have caught on with other writers.

Collecting evidence using different types of assessment over time aids

validity, but raises reliability concerns since the measurement circumstances are

varied. Consequently, classroom assessment creates an interesting paradox where

validity and reliability appear to be opposing forces (Black and Wiliam, 2012).

Fortunately, if the assessments reflect the content standards they do lend

themselves to dependability, thus balancing reliability and validity. Furthermore,

professional judgement could be used to watch for unreliability and invalidity in

individual assessments (Harlen, 2005a). Herbst and Davies (2014) believed that informed professional judgement could be considered reliable and valid if done consistently and reflects the set standards. For instance, the use of rubrics helps to demonstrate how a grade was determined (Harlen, 2005b). However, how likely is it that a classroom teacher can pinpoint an inclusive grade on a 100-point scale? How can we believe that a grade determined by one educator will be the exact same as by another across the province if shown the same evidence? Quantitative studies usually include a stated margin of error, but there is no such information on a school report card. Participants are literally expected to take an individual student's most consistent achievement, with special consideration for more recent achievement, and decide upon a percentage grade. It is not possible to make such judgements with absolute accuracy, yet grades are still held sacrosanct. However, this fact appears to be sheltered within shadowed practices. Even admitting a small margin of error could be deemed to indicate a flawed system and lead to public repudiation, so assessment remains an enigmatic process (Wilson, 1998; Morrison and Wylie, 1999; Newton, 2005).

Others have also explored the place of validity in classroom assessment in various literature reviews and discussions. For example, Crooks, Kane and Cohen (1996) pointed out numerous factors incorporated in assessment and argued that validity is only as strong as its weakest link. Shepard (1993) and Wiggins (1993) advocated discussing assessment validity to explore how to fix design issues and interpret assessment properly. For instance, assessment should be logically connected to the intended content, i.e., content validity (Frisbie and Waltman,

1992; Popham, 2009). Likewise, Airasian, Englemann and Gallagher (2007) stated that assessment and instruction must be conceptualized as one educational component. Participants in this study suggested they attempt to be consistent and accurate in their practices. They also stated they would not hesitate to defend their assessment practices to students and parents, but the willingness to stand by their professional judgement does not in itself remove the threats to reliability and validity evident in Heuristic Assessment. Again, Heuristic Assessment has a beneficial proclivity towards students; erring on the side of caution is rarely questioned. At the same time, if we can address the flaws associated with Heuristic Assessment, we can propose a better assessment model.

4.7.1 Construct Validity, Construct Irrelevance, and Social Consequences

Over the last few decades, the "construct validity" idea has achieved consensus as being the most appropriate approach to assessment. Cronbach (1989) helped to start the conversation on construct validity as a means of joining reliability and validity. The idea was further advocated by researchers such as Messick (1989a; 1989b), Shepard (1993), Airasian (1994), Moss (1992; 1994; 1995), James (1998), Newton (2005; 2012). Messick (1995, p.5-6) argued that it is how assessment is interpreted and acted upon that established validity. He described "six distinguishable validity aspects… content, substantive, structural, generalizability, extended, and consequential aspects." In other words, construct validity provides the means for an educator to take evidence of learning, confirms the situation allowed for students to display their achievement on multiple occasions, reflects

how this achievement should be interpreted, and uses the information to express achievement as a grade.

In *Growing Success* the definition of validity and of professional judgement reflect the influence of the concept of construct validity. Standards-based educational reform produced content standards as well as performance standards. Construct validity is established by combining an assessment task with the content standards. Furthermore, the teacher is asked to reflect on this content and criteria, as well as context, evidence of learning, means of instruction and assessment, and combine all it all as a whole. Professional judgement has its own sense of reliability and validity (Sutton, 1991; Davies, 2000; 2007; 2011) and can be consistent with the parameters of construct validity (Shepard, 2000b). Teachers need such autonomy in order to do their jobs (Daugherty, et al., 2012). Evaluation becomes less quantitative and begins to mimic how qualitative researchers look at evidence and reach conclusions (Moss, 1992; Gipps, 1994). Ensuring construct validity is the foundation of good professional judgement. However, Heuristic Assessment, introducing variables that can be unique to a particular student, is initiated as a result of the situation, but can be a threat to construct validity.

When policy said that an educator must consider context when conducting an assessment, this direction may be confused for considering situational factors instead of achievement. If teachers include non-curriculum elements within an evaluation, or are inconsistent in the way they assess, the result is construct irrelevance (Messick, 1989a; 1989b). It is perhaps the greatest threat to a "valid" evaluation (Anders and Richardson, 1992; Brookhart, 1993; 1994; Cross and

257

Frary, 1999; Friedman and Troug, 1998; Moss, 2003; Stobart, 2008). In the interest of helping students, the participants displayed numerous examples of this hazard. Also, if an administration attempts to raise reliability by forcing teachers to match achievement to overly specific criteria, or to follow universal calculation rules, this can actually belie evaluation (Newton and Meadows, 2011). Sutton (1991) wrote that professional judgement could be used to decrease elements that interfere with accurate assessments, but could not eliminate all factors. The goal is to make informed decisions in a timely appropriate manner. Again, educators must be aware that assessment is ultimately subjective, and there is no panacea for measurement error (Cooper, 2007; Koretz, 2008). However, by exercising construct validity, and having awareness for construct irrelevance, the teacher can establish a trustworthy evaluation.

Heuristic Assessment can also be influenced by another factor of validity first identified by Messick (1989a): social consequences. Messick changed the way classroom assessment was viewed when he proposed that social consequences, the resulting inferences and aftermath of an evaluation, were an element of evaluation validity. Again, Heuristic Assessment errs on the side of caution, which usually results in assigning the highest justifiable grade. As a result, the chances of good social consequences increase. For example, the participants spoke of post-secondary acceptance as well as scholarships for their students. Nevertheless, construct irrelevance is at play, even if it is done with good intentions. It is another dilemma that needs to be explored.

Assessment experts tend to agree that social consequences, i.e., how grades

are postulated, are a significant part of evaluation (Seligman, 1998; Stiggins, 2001b). Newton (2005) pointed out that it is also an ethical consideration: grades cannot be increased on the basis that the student is a "good kid" and "deserves" a higher mark to improve future life options, as there is such criteria is completely subjective. Since a reflection on social consequences can skew an objective measurement, there is debate whether social consequences should be part of the grade gestalt, as described by Messick (1989a; 1989b), or are a separate entity, as proposed by Shepard (1997). Either way, how should an educator best account for social consequences? The matter is further complicated when the teacher reflects on how the students may have felt when they were being assessed, such as test anxiety. If students were truly intimidated by a summative assessment, and this feeling interfered with displaying their best evidence of learning, should this be a validity issue (Brookhart, 1991; Gardner, et al., 2010)? The respondents talked about getting "to know" a student over time, suggesting that they could see evidence of achievement that may not have necessarily come through on a given day or evaluation. On the other hand, concern for social consequences could be dismissed as an exercise of second guessing an otherwise valid assessment. However, the main point is validity in assessment is a range of variance, and not a black-or-white issue (Messick, 1994; 1995; Kane, 2001). Fleer (2015) demonstrated in a study of eleven Australian primary teachers that students experience the assessment situation in different ways where background and sociocultural issues plays an important and relatively unexplored role. Informed professional judgement is needed to help educators navigate this range of variance and determine a dependable overall score that balances all applicable factors of

achievement.

Evaluation takes place in a complex social process (Mehan, 1998; Wenger, 1998; Smith, 2003). No one evaluation is valid for all test takers (Sutton, 1991; Newton, 2007). A teacher can base an evaluation on content standards, with a well-designed scoring rubric and clearly articulated instructions. Nevertheless, given the ultimately subjective nature of assessment – there will always be flaws relative to certain students. Therefore, social consequences will always be need to be considered, either an aspect of validity or a separate byproduct. Educators need to consider the situation and multiple angles to interpret appropriate meaning, which includes gathering as much evidence of achievement as possible (Cherryholmes, 1988; Messick, 1989a; 1989b; Cronbach, 1989; Tittle, 1989; Moss, 2003). However, the ability to collect evidence can be interrupted by a multitude of issues (i.e., students not participating in summative assessments by choice and/or outside factors), and a teacher does not have the omniscience to foresee how a particular evaluation was not appropriate for an individual student. This reality helps to explain why Heuristic Assessment happens in the first place, but does not mean that assessment practices cannot be further improved.

4.7.2 Equity, Fairness, and Transparency in Assessment

Social consequences can also be understood in relation to equity and fairness, two terms that have come up frequently in this study. In one *Growing Success* source, Volante (2006) did not discuss professional judgement per se, but he did make a key observation about reliability and validity in grading. To ensure grades are equitable and fair, educators must pay attention to the validity of

assessments. Again, the relationship between reliability and validity is

problematic in student evaluation, but if the teacher focuses on construct validity

where reliability is treated as a facet, equity and fairness can be realized (Moss

1994; 1995; Whittington, 1999). However, ensuring equity and fairness is not an

easy task, as assessment is always built on subjective ground.

When designing assessments, educators need to consider what they know

about the background of students (Baker and O'Neil, 1994; Garcia, 1995; Gipps,

1995; Stobart, 2005). In a multicultural province such as Ontario, this is a definite

challenge. However, such genuine reflection adds to validity. Should an educator

ignore signs that instruction was not understood because, for example, analogies in

lessons have been based on cultural material unfamiliar to students, then an

evaluation of this material loses fairness and equity. Moreover, evaluation would

lose validity while adding to the likelihood of negative social consequences. At

the same time, a teacher who realizes the flaws in the individual evaluations too

late may then apply Heuristic Assessment to raise a student's overall grade. In

other words, the adjustment is being done out of a sincere desire to be fair and

equitable to the student and limit social consequences, but if the adjustment is not

based on actual achievement – this is not the intention of professional judgement.

Earl, et al. (2010) suggested that professional judgement must be accurate,

fair, and equitable, but did not further elaborate. Rowe and Hill (1996), in their

study of Australian student profiles, as well as the findings in Harlen (2005b),

asserted that educators must make evaluation criteria clear in order to ensure

fairness and equity. After all, researchers tend to agree that students, above all,

expect their teachers to be fair (Brookhart, 1993; Joint Committee, 1993; Hargreaves, 2004; Guskey, 2004; 2006; Western, 2006). Making criteria clear also relates to the other key fundamental principle of transparency. However, this study has cited frequent examples of the unexpected conflicts of trying to adequately combine the fundamental principles of equity, fairness, and transparency.

Throughout this study, the concept of shadowed practices has been discussed. A simple definition for this term would refer to assessment actions taken by teachers, often at the behest of the local administration, that are kept secret from students and other stakeholders. However, *Growing Success* suggested that assessment and evaluation should be transparent. Technically, this guideline could only refer to the use of rubrics or other devices that inform students of what is expected on a summative assessment and how it will be evaluated. Still, the idea of transparency suggests that, without it, grades are attenuated. The literature showed that the issue of a lack of evaluative transparency is not limited to what is described by the participants in this study (Cumming, et al., 2006; Wyatt-Smith, Klenowski and Gunn, 2010).

It appears *Growing Success* borrowed from the Joint Committee (1993), as well as Harlen and Deakin Crick (2003), when they emphasized the need for transparency in practices for students and parents/guardians as part of a school's culture. This position is supported by others assessment authors, such as Gardner, et al., (2010). However, true transparency is difficult to accomplish. Assessment is complicated. Trying to explain the intricacies of determining a student's most

consistent / more recent achievement, even when a teacher is doing everything properly, is difficult to explain to a general audience. Newton (2005) believed that, as soon as an instructor starts to talk about the "inaccuracy" of assessment and evaluation – some would be led to the assumption that something is not being done correctly. Consequently, educators prevaricate when discussing evaluation with stakeholders, ultimately benefiting no one.

Social values are important and are an aspect of achievement, but they are much more challenging to articulate (Sadler, 1987; O'Donovan, Price and Rust, 2004; Stake, 2004; Wyatt-Smith, Klenowski and Gunn, 2010). When providing advice to educators on how to create an evaluation plan, Frisbie and Waltman (1992) wrote that the point of grades is to communicate achievement. Therefore, to knowingly interfere with transparency defeats the purpose. Furthermore, Gardner (2012b) believed that the expectation of transparency could counter-intuitively encourage dishonesty. As educators need to demonstrate accountability, they may actually cover up actual actions with more socially accepted ones (Stiggins, 2005; Koretz, 2008; Hayward, 2015). This idea was demonstrated in the analysis, such as the shadowed practice of promotion meetings. The intricacy of professional judgement, at least in the form of Heuristic Assessment, can diminish transparency instead of improving it. It is not that the participants in this study desired to mislead stakeholders for nefarious purposes; they just want to simplify a problematic process by presenting one that is easier to explain and understand.

Ontario has attempted to provide clear standards while allowing room to

suit individual circumstances. At the same time, the OME hesitates to directly

address the realities of evaluation. There is no use in hiding the limits of

classroom assessment validity. Furthermore, no set of standards can be general or

specific enough for the real world (Sadler, 1987). How to consider outside factors

should be clearer to educators (Moss, 2003; 2013). The public could also learn

more about the role of social consequences so it can better understand assessment

(Harlen, 1994; Earl, 1995; Davies, 2000; 2007; 2011; Popham, 2003; Newton,

2005; Cooper, 2011). Unfortunately, establishing better transparency is held back

out of the fear of a petulant public, and so the cycle perpetuates (Black, 2003).

4.7.3 Professional Judgement and Adjustments

This discussion brings us to what could be called the modus operandi of

Heuristic Assessment: adjustments. Adjustments, or the metonymic "bump," were

discussed throughout this report's third chapter (e.g., Sections 3.2.4, 3.4.3, and

3.5.3). They could be described as supervening professional judgement. Instead

of using informed professional judgement as part of triangulating student products,

conversations, and observations, with special consideration for more recent

achievement, an adjustment is made when a percentage grade is determined, then

modified. The analysis illustrated that adjustments permeate the system, even

though there is no reference to them within *Growing Success*. There is a plethora

of reasons for adjustments, with most having to do with what assessment experts

would refer to as construct irrelevance. The justification, based on the voices of

the respondents, suggested adjustments were done based on situational elements

and a reflection on social consequences. Besides, adjusting a grade by a few

percentage points was seen as having far more benefits than doing any actual harm. However, our goal should be to eliminate shadowed practices, and to show educators there is way to achieve similar results that satisfies construct validity.

In line with social consequences, researchers have noted that students take grading personally (Harlen and Deakin Crick, 2003; Harlen, 2006; Musial, et al., 2009; Ross and Kostuch, 2011). Because of the cultural currency assigned to grades, adjustments are influenced by concepts such as fairness. Like the participants in this study, Wyatt-Smith, Klenowski and Gunn (2010, p.68) found that the teachers in their study were likely to give students "the benefit of the doubt." In a follow-up study to Joint (1993), Scott et al (2013) surveyed 3312 Albertan stakeholders, from teachers to students. The findings show an expectation from all parties that behaviour is part of grades and good students should be rewarded. Although increasing grades is done out of good intentions, validity problems occur (Frary, Cross and Weber, 1993; Grace, 1993; Gullickson, 1993; Plake and Impara, 1993; 1997; Cizek, Fitzgerald and Rancor, 1995; Shepard, 1997; 2000a). In addition, when teachers begin to "bump" individual student grades, they may begin to look at the class average. In order to be fair, if one student's grade is to be adjusted, should it be done for the whole class (Airasian, Engemann and Gallagher, 2007)? Adjusting the grade for one student can cause a domino effect.

Gardner, et al., (2010) suggested that teachers may also increase marks when they feel they are being watched, as higher grades are actually less likely to be challenged by stakeholders (Pollio and Humphreys, 1988; Harris and Brown,

2009; Zhu and Urhahne, 2015). Classroom assessment can be used for different purposes, especially when reflecting on the desires of particular stakeholders. As was noted in the analysis, participants felt conflicted between assigning a grade that was truly earned, and both explicit and implicit expectations by third parties. Consequently, grades become a putative product instead of a reflection on an integrative learning process, with the end goal of maximizing the quantity over quality (Newton, 2012; Pugsley, 2012).

The makers of the evaluation program *Markbook* were contacted regarding a feature in their software that automatically rounds up a nine (described as "Bump 9s."). When asked about the reason for this option, a representative said it allowed teachers to "feel better about [the overall grade]" (Chellew, 2013). The representative also made reference to the emotional reality of assessment. The common perception that a 1% increase does not distort actual earned achievement, but can make both student and teacher "feel better," translates to participants being in favour of small inflations to percentage grades – especially if it improves post-secondary opportunities. Being sensitive to the feelings of students, along with uncertainty of the good grading practices, leads to erring on the side of caution (Barnes, 1985). However, the analysis made the argument that 1% does not always equal 1% in regards to classroom assessment (see Section 3.5.3). Rounding up by 1% usually means modifying the entire level of overall achievement (Gardner, et al., 2010; Black and Wiliam, 2012). For example, a 79% is a 3+, or consistent achievement, which *meets* the provincial standard; but 80% is 4-, or consistently *exceeding* the provincial standard. Furthermore, there is

the whole other concern of how Ontario teachers must translate an overall level of achievement into a percentage grade, which also contributes to the issue of adjustments (Cooper, 2007; McMillan, Hellsten, and Klinger, 2011).

Although the participants kept detailed records for the summative assessment of student products, they admitted to relying on memory when it came to conversations and observations. However, memory alone is not a valid form of record keeping, especially when educators confound the achievement of different students (Linn and Gronlund, 1995; Miller and Linn, 2000; Trouilloud, et al., 2002). In a survey by Stiggins and Bridgeford (1985), 228 American educators representing all grade levels were consulted. These participants also confirmed a reliance on mental record keeping, which contributed to invalidity and unreliability in overall grades. On a similar note, Hay and MacDonald (2008) applied a Charmaz constructivist grounded theory approach to a study of Australian physical education teachers in Queensland. They found that the teachers applied professional judgement based not on state standards, but their own memory instead of an actual compendium. Consequently, impressions of students, particularly those who applied themselves and made themselves known, benefited more than shy students. This approach created construct irrelevance in grades. Even though the Ontario assessment system is meant to use most consistent achievement, what an educator believes he or she remembers about a student's achievement, as demonstrated in the analysis under the section Perceptions of Student Behaviour, could interfere with pinpointing a percentage grade.

4.7.3.1 Hodgepodge Grading

Hodgepodge grading was a term coined by Brookhart (1991). It is a general description given to the gallimaufry of evaluation practices that are clearly not based on achievement factors. Mixing student achievement and behaviour has been noted in the literature since the 1950s and is a prime example of construct irrelevance (Frary, Cross and Weber, 1993; Cross and Frary, 1999). For instance, Allal (1988) found no consistency in how her participants determined grades, e.g., different procedures for different subjects, no uniformity in adjustments and calculations, mixing norm and criterion referencing, etc. Allal also suggested that grades were being used as a form of classroom management. There was clear evidence of hodgepodge grading in the analysis. Like most shadowed practices, hodgepodge grading begins with the idea that "good" students should be rewarded. Furthermore, it is seen as a way to relieve some of the social consequences of grading by justifying an adjustment to an earned grade. However, since these adjustments are mostly based on non-achievement factors, the overall grade becomes spurious.

After Brookhart (1993; 1994) expanded on her ideas by illustrating her ideas using a study of 84 American educators, other researchers investigated hodgepodge grading in other education systems. For example, Cizek, Fitzgerald and Richer (1995) survey 60 American middle and high school teachers and instantiated non-achievement factors prevalent in marking. Cross and Frary (1999) also found that participants had a penchant for non-achievement factors. Furthermore, students and other stakeholders felt the approach was a valid way to

reward good work. Likewise, McMillan (2001) examined 1483 Grade 6-12

teachers within 53 urban Virginia schools and found similar results. In a follow-

up study of twelve different English teachers from twelve schools in seven

American districts, McMillan (2003) found further confirmation of hodgepodge

grading. Studies by Hoge and Coladarci (1989), Stiggins, Frisbie and Griswold

(1989), Plake and Impara (1993), Joint Committee (1993), Tittle (1994), Marzano

(2000), McMillan and Nash (2000), Shepard (2000a), Guskey and Bailey (2001),

Guskey (2004), Airasian, Enngelman and Gallagher (2007), Duncan and Noonan

(2007), Musial, et al., (2009), Black and Wiliam (2012) also corroborated such

findings.

Earl, et al., (2010), similar to this study, also found the practice evident

Ontario schools. In their text, McMillan, Hellsten and Klinger (2011) wrote that

hodgepodge grading created a dual system in Canadian schools: educators look for

signs of improvement and criterion-referencing to evaluate weaker students,

whereas strong students are evaluated for actual overall achievement based upon

normative-referencing. Consequently, the meaning of grades becomes blurred,

and it is difficult to winnow out a student's true score. Numerous other writers

reached a similar conclusion, stating that hodgepodge grading made report cards

more difficult to interpret by other educators and stakeholders (Gronlund and Linn,

1990; Friedman and Manley, 1992; Friedman and Frisbie, 1995; Guskey and

Bailey, 2010).

In light of the evidence of hodgepodge grading, assessment experts such as

Gronlund and Linn (1990), Ebel and Frisbie (1991), Friedman and Manley (1992),

and Airasian (1996) have urged educators to consider only achievement factors for grading. Unfortunately, correcting the situation has proved difficult. Larson (2009), in her survey and interviews with 125 Ontario educators pointed out that proper assessment techniques are included in new teacher initiatives in the province, but they rarely fulfilled as the reality of the classroom sets in. This observation could also be noted in the analysis. Consequently, grades become variegated and lose consistency.

There are several possible explanations for why hodgepodge grading occurs. Earl, et al., (2010) stated that half their participants were pressured by rapacious students and/or the administration to raise grades. Furthermore, 22% of the participants claimed that their administration unilaterally raised grades. These two points are consistent with the findings in this study. Again, it is possible that stakeholders actually expect a degree of hodgepodge evaluation, based on their misunderstanding of what grades are meant to convey. Teachers may also resist "proper" measurement as educators evaluate the work of students on the basis of how they perceive they were marked as students (Guskey, 2006). Furthermore, Allal (1988) and Brookhart (1994) thought hodgepodge grading discouraged egregious classroom behaviour.

4.7.3.1.1 Influence of Student Behaviour on Grades

Although hodgepodge grading could reflect any non-achievement factor that is included in a grade, the perception of how a student behaves appears to be the main motivator. In this study, there was an entire conceptual category (see

Section 3.4 Perceptions of Student Behaviour) to show its impact on Heuristic Assessment. The study shows that behavioural factors include attendance, effort, and punctuality. Even when students appeared to be acting in a certain way just to ingratiate themselves to the teacher, studies suggested instructors would prefer the façade of effort as opposed to unacceptable behaviour.

Effort, in the form of several epithets, was frequently cited in this study as something that was important to the participants. Stiggins, Frisbie and Griswold (1989) found almost all their participants considered effort when grading. Friedman and Manley (1992) asked similar questions to 227 participants, comprised of teachers, students, administration and counselors, from 35 different Wisconsin High Schools with comparable results: effort was valued along with actual achievement, particularly by students themselves. When students are advocating for effort to be included in grades, its importance in the situation is emphasized. Cross and Frary (1999) found 74% of student supported such hodgepodge methods of assessment – especially because students felt they benefited from such an approach. In another study, Howley, Kusimo and Parrot (2001) interviewed 52 female American Seventh Graders and their teachers and also found effort to be a significant part of the grading process.

Educators do need to reflect on student effort as part of assessment as and for learning, but the challenge is not to conflate effort with assessment of learning (Brookhart, 1997; McMillan, 2008). As Stiggins (1988; 2005) pointed out, students can pretend to work hard if they know there are marks for effort. Nevertheless, Stiggins and Conklin (1992) claimed in their text that teachers tend

271

to consider effort indiscriminately in their grading practices. It is not surprising, since school is meant to prepare students for real life, and the general belief is life rewards effort. Stiggins (2005) believed educators should encourage effort, but because it cannot be measured mathematically it has no place in a grade based on achievement. Kohn (1993) also disagreed with the notion that students who "go the extra mile" deserved a positive adjustment, but this reasoning is often lost in the evaluation process. For instance, Cross and Frary (1999) stated that 25% of their participants made significant adjustment based on effort. Some the participants in this study suggested a grade could change by a largesse 5%, mainly based on the perception of effort and other positive behaviours.

Educators may feel that they are benefiting good students when effort is included in a grade, when they are actually grading the learning process itself. Instead of encouraging intellectual development, in an environment where everything is summative, creativity can be stifled. Another important point raised in Howley, Kusimo and Parrot (2001) was when students consciously acted within a matrix of a certain kind of behaviour, it reduced academic risk taking, a finding that was also noted by Aaronsohn, et al., (1994) in their interviews with American elementary, middle, high school, and university students. In other words, when learning is seen as a product, and the goal is to get the highest grade possible, students will conform to what are perceived to be tried-and-true steps instead of being imaginative. The possibility of falling short and displeasing the teacher could result in a myopic view of what they could accomplish. In this situation, grades actually get in the way of the learning process. Students may feel that the

grades they receive accurately depict a fixed level of achievement. This would help explain why the participants of this study indicated that students tended to achieve at the same level throughout secondary school (see Section 3.5.3). Klapp's (2015) longitudinal study of 8558 Swedish student found that the final grades in late elementary school stayed mostly consistent throughout secondary, with struggling elementary students the most likely not to finish secondary.

Teachers can also interfere with the reliability and validity of their own grades when they include behavioural aspects (McMillan, Hellsten and Klinger, 2011). Frary, Cross and Weber (1993) found the daily performance of the student, including attendance, homework, and participation were seen as a way to fairly determine overall grades. Such assessment practices were also noted by Messick (1989a), Friedman and Manley (1992), Zeidner (1992), Airasian and Jones (1993), Matanin and Tannehill (1994), Cross and Frary (1999), Wyatt-Smith (1999), Howley, Kusimo and Parrot (2001), Stiggins, (2001b), Airasian, Engemann and Gallagher (2007), Matteucci, et al., (2008), and Newton (2012). Hay and MacDonald (2008, p.11) also found that their participants applied professional judgement based not on state standards, but their own interpretation of what they believed was appropriate. Just as participants in this study referred to "knowing" the student and saying it was "unrealistic not to include [behaviour]," Hay and MacDonald reported that their participants referred to "gut feelings" and that one "just can't avoid" including the perception of a student's behaviour within a grade because it is "locked in your brain." Their respondents also had certain grades in mind before they reflected on the content standards, further adding to construct

irrelevance. The participants in this study also suggested the impression of students could have an ineluctable influence on adjustments.

This study suggested that the participants usually strived to justify the highest grade possible. McMillan, Hellsten and Klinger (2011) reached a similar conclusion, also supporting the influence of student behaviour on the decision. Airasian, Engemann and Gallagher (2007), as well as Black and Wiliam (2012), said it is understandable that educators want to be generous with grades, but they need to focus on actual achievement and not perception. Furthermore, between the responses from the participants, as well as the correspondence with the representative from *Markbook*, there is a belief within the assessment arena that erring on the side of caution makes teachers feel better about the grades they assign. Moreover, increased grades decrease the likelihood of having to justify decisions to stakeholders (Stiggins, Frisbie and Griswold, 1989; Stiggins, 2005).

Fortunately, there was no evidence that the participants in this study reduced grades because of undesirable behaviours imputed to students. However, such students were unlikely to get a positive grade increase, except in cases of borderline students. Some researchers have discovered a similar pattern in other school systems (Stiggins, Frisbie and Griswold, 1989; Matanin and Tannehill, 1994; Wyatt-Smith, Klenowski and Gunn, 2010). Cross and Frary (1999) stated most of their participants did not lower marks based on a lack of effort, even though many participants increased marks based on overt effort. Surprisingly, 81% of teachers and 70% of students in their study approved of a separate mark for behaviours such as effort, despite the general finding that effort should be

included in the grade. If these numbers are accurate, it is good Ontario has a

separate report card section for learning skills, where teachers can make a

summative judgement on behaviours.

Nevertheless, educators can have unconscious predilections in their

evaluation practices. Teacher bias is also a form of hodgepodge grading as it uses

non-achievement factors to contribute towards the overall grade. For example, the

impression a student's achievement makes on a teacher can also lead to the halo

effect. When an educator comes to expect a certain level of achievement from a

student, they may evaluate an assignment with a certain level in mind and mark

accordingly. This phenomenon has been noticed by Nitko (2001), Harlen (2005b),

Wyatt-Smith and Castleton (2005), Airasian, Engemann and Gallagher (2007), and

Cooksey, Freebody and Wyatt-Smith (2007).

4.7.3.1.2 Using Formative Evidence for Summative Purposes

Formative assessment and summative assessment are not a dichotomy; the

difference between the two is how evidence of learning is used. If evidence is

used as part of the learning process, it is formative; when evidence be used to rate

a student's achievement, it is summative. In the interest of transparency, Ontario

students are to be informed when as assessment is summative, i.e., an evaluation

or assessment of learning. However, in the grey areas of the situation, students

demonstrate learning in different ways that do not always fit effectively into one of

two categories. A particular summative assessment might not have an adequate

design to allow a student to show knowledge and ability. Since an Ontario

student's grade is to be based on most consistent achievement with special consideration for more recent achievement, ideally, an educator has time to collect multiple student samples to validate a grade. However, the participants in this study would testify that the ideal is too simplistic. When it comes time to record the final report card, how does more recent achievement tip the scales in a way that satisfies the fundamental principles? When triangulating student products with conversations and observations, does evidence collected during formative circumstances automatically become summative or is a case-by-case basis? If an educator can recall formative evidence that would benefit the student, would giving a positive adjustment be an example of hodgepodge grading, teacher bias – or informed professional judgement? The magnitude of such questions quickly replaces an ideal approach with Heuristic Assessment.

As we read in the Introduction of this study, *Growing Success* hinted that a formative evidence of learning, including classroom observations and conversations, is fungible with summative evidence – based on a teacher's professional judgement – when determining a final grade. However, emphasis appeared to be placed on student products designed for evaluation when triangulating overall achievement. Sutton (1991) claimed that conferencing with a student can provide excellent evidence of learning, but admitted it is difficult to do regularly. Furthermore, it is problematic to keep detailed records of student products, conversations, and observations for formative purposes, and another set of three for summative purposes. Consequently, participants in this study relied on memory of student achievement to some degree. After all, the perception of

students created its own sense of consistency. When this impression was compared with the summative records of student products, Heuristic Assessment then instructed the participants of whether or not an adjustment was needed. This action may sound like triangulation, but it is not true informed professional judgement.

In short, the participants felt that the "hard working" students should have their efforts recognized. As a result, the ones who did their homework and participated in class were likely to receive a positive adjustment of the overall grade, even though student behaviour should not be part of the equation. The justification is not only do these students demonstrate a desire to learn, but make the teacher's job easier. The literature has noted teachers lamenting many students apply themselves only when they know an assignment "counts" for marks (Crooks, 1988; Sadler, 1998). In the Earl, et al., (2010) study, they found that 26% of Ontario teachers included homework as part of final mark. Likewise, the participant in the Simon et al. study (2010) told her students that homework would count in order to motivate them to do it. Cizek, Fitzgerald and Rachor (1996) found it was common for formative evidence to be used for summative purposes. This claim follows the general finding of this study: educators act in a way that makes sense to them with the objective of improving student learning. Informed professional judgement should allow an educator to reflect on evidence collected during formative assessment as a comparison to the consistency collected during summative and weighted appropriately during triangulation. However, the sense from the participants is that Heuristic Assessment is not guiding them to reflect on

the formative assessment per se, but the positive and negative behaviours of the students, and to adjust accordingly. This shadowed practice lacks construct validity.

Mixing formative and summative evidence in an overall grade once again brings up the issue of reliability and validity. Harlen and James (1997) warned that confusion between formative and summative assessment could damage the reliability and validity of both. Assuming that an educator understands the difference, Guskey (1994) hinted that process and progress could be included in a grade, but Gipps (1994) and Cross and Frary (1999) disagreed. Cooper (2007) stated that a teacher must aim for adequate summative evidence to avoid the debate altogether, but we have seen in the situation there are challenges when students do not complete all assignments. If using formative evidence is done with consistency, then there is an argument for reliability (Gardner, et al., 2010). Again, if the formative and summative assessment point to similar conclusions, then they could be used to support one another (Speck, 1998; Stiggins, 2005). After all, how can any evidence of learning be ignored (Smith, 2003)? On the other hand, the daily classroom must make a distinction between formative and summative assessment to aid the learning process. Harlen and Deakin Crick (2003) pointed out that if students think everything is summative they will be more afraid of making mistakes, hence limiting risk taking and threatening imagination. Harlen (2005a) suggested that formative assessment could be used for reliability, and summative for validity, as an effective way they can both contribute towards evaluation. This view is what Stiggins (1997) might have called sufficient depth,

or Smith (2003), sufficiency of information. In the Ontario assessment situation, a good framework has been established to encourage formative assessment in the everyday classroom to aid the learning process, with a course designed to allow for the collection of multiple samples of summative assessment to base an overall grade. The writings of Black (2015) as well as Elwood and Murphy (2015) argued that debating the differences between formative and summative assessment is not moving the conversation forward on how assessment data should work together. Educators must not be hesitant to actively gather observations and have conversations with student to augment student products and inform their professional judgement. When professional judgement is active, most consistent achievement, with special consideration for more recent achievement, evaluation of a final grade will seem less like a last step and more like the fruition of the learning process back by consistent and transparent evidence (Gill, 2013b). When informed professional judgement is applied, formative and summative assessment are not two sides of the same coin, but two points on the same line.

4.7.4 Debating Punitive Measures

As previously mentioned in Section 3.2, when addressing professional judgement and the final report card, it is difficult to ignore punitive measures or negative adjustments to a student's grade. Because applying a punitive measure is such a contentious issue (see Section 1.6.5.5), and such measures are applied to individual assignments, we must be careful not to lose focus on the main research topic. In short, we saw that the participants supported punitive measures as a last resort, on the condition they were not mandatory and seen only as an option to

spark motivation. This position was mostly based on a conceptualization of what happens in the real world when someone does not do his or her work. Some participants claimed it worked, and any mark deductions were limited to avoid any deleterious effects. Still, we saw that most of the participants continued using average when calculating overall grades. Consequently, punitive measures were more likely to hinder determining a student's overall achievement with accuracy, which goes against what *Growing Success* said about the issue.

In the literature, the vast majority of assessment writers disapprove of punitive measures (Frisbie and Waltman, 1992; Selby and Murphy, 1992; Kohn, 1993; 1996; Brookhart, 1999; Costello and McKellar, 2000; Reeves, 2004; 2008; 2010; Western, 2006; Airasian, Engemann and Gallagher, 2007; Cooper, 2007; 2011; O'Connor, 2007; 2009; McMillan, 2008; Cooper, O'Connor and Wakeman, 2009; Musial, et al., 2009; O'Connor and Wormeli, 2011; Davies and Herbst, 2014). The general argument against punitive measures is that they distort calculations of true achievement, are more likely to demotivate than to encourage future learning, and are inappropriately used in comparison to "real world" scenarios.

The main argument in favour of punitive measures is they are deemed to mimic the future realities of responsibilities in the workplace (Steffenhagen, 2010). Educator support for punitive measures was examined by Brookhart (1993), who found half of the respondents she consulted felt missing work should get a zero, even if it meant failing. The other half still supported zeros, but only down to a passing grade. Similar results were found in a follow-up study

(Brookhart, 1997). McMillan and Nash (2000) also noted that zeros were seen as a motivator, be it for formative or summative assessment. Furthermore, they confirmed the general belief that graded work was more likely to be completed by students. Although most assessment writers are against punitive measures, Stiggins (2005) agreed that punitive measures could have some benefits, such as correcting negative student behaviour. Earlier, he wrote that punitive measures could teach life lessons about showing up to work on time and completing work (1992). However, he said punitive measures should not take away from what a student does accomplish. Also, students should be given extra chances. He preferred that punitive measures not be applied to the grade, instead, students should be required to complete work in order to participate in rewarding activities such as field trips. Other assessment writers agree that punitive measures should not be used in a threatening manner (Canady and Hotchkiss, 1989; Stiggins and Duke, 1991; Guskey and Bailey, 2001; Airasian, Engemann and Gallagher, 2007; Reeves, 2010).

Respondents in this study repeated the above sentiments, as they felt they needed punitive measures within their repertoire of assessment tactics in order to engage all learners (see Section 3.3). Zwaagstra (2012a; 2012b) supported this position on the grounds that having no punitive measures can interfere with the professional judgement as it removes a viable option. Earl, et al., (2010) found Ontario educators struggling with cheating and plagiarism. Even though punitive measures were under a moratorium, 83% of the participants admitted to using some form of mark deduction including the use of zero. Although 78% were

allowing replacement assignments with no penalty, 80% said they would include

zeros within the final mark calculations. These numbers appear consistent with

what the participants in the present study reported. The most important point here

is that even if a teacher does apply punitive measures to an individual assignment,

informed professional judgement must revisit such decisions as part of the final

report card process. Zeros should not be included in the calculation if they will

distort the true determination of overall achievement and more appropriate

evidence is available. This is an instance where evidence obtained from formative

circumstances will not simply give a student a higher grade for the sake of it, but

the higher grade will have more construct validity as it better reflects most

consistent achievement.

4.7.5 Passing Borderline Students

If students have punitive measures such as zeros and late marks on their

records, they may also be in peril of failing the course (see Section 1.6.5.4.4). If

the grade as reported by software puts them in the forty percentile, the classroom

teacher will need to decide whether or not to increase the student to 50% based on

a paucity of achievement evidence. The analysis demonstrated that, when faced

with this situation, all the participants would increase the grade with the approval,

or direction, of the local administration (see Sections 3.3.5 and 3.4.3). The

literature also suggested that there is an aversion to failing students across

educational systems. Moreover, the decision to pass the student is usually based

on non-achievement factors as much as achievement. Therefore, borderline

passing grades are often examples of construct irrelevance.

Frisbie and Waltman (1992) pondered what a student needs to do to pass with reliability and validity, and proposed that it should be based on a consensus with respect to certain standards. Terwilliger (1989) and McMillan, Hellsten and Klinger (2011) said it was acceptable to pass borderline students – as long as the decision is based on minimum achievement to satisfy construct validity. However, Stiggins, Frisbie and Griswold (1989) found, in reality, non-achievement factors play a significant role in whether or not a borderline student passes. Likewise, Brookhart (1993) stated the exculpatory evidence to justify passing such a pupil is overt effort. Furthermore, she pointed out that teachers, like the participants in this study, would rarely lower a student's grade to a failing level. Gronlund (1998) claimed that passing borderline students was not detrimental to their development. Cooper (2007) disagreed, believing that this approach to passing weak students is a reverberation from normative-referencing. He also felt that passing borderline students is only setting students them up to fail in the next grade. Indeed, the participants noted that if a student is borderline, s/he is likely to be borderline in the next year/subject. The shadowed goal appears to be to continue to push the student along to graduation. As a result, the OME can demonstrate statistics of ever increasing graduation rates, thus improving public confidence in the system (*MacLeans*, 2011; Ontario, 2010c; 2013a; 2014; Office, 2015; Rushowy, 2015). The public may not understand construct validity, but having the vast majority of students passing appears to be a good return on taxpayers' investment.

Failing a student not only has an emotional impact on a student, but the educator as well. Passing a borderline student can also make a teacher feel better

about the situation (Wyatt-Smith, Klenowski and Gunn, 2010). Airasian (1996) and Airasian, Engemann and Gallaher (2007) claimed that passing a borderline student provided relief to both the student and educator, even if the decision creates construct irrelevance. After all, it was argued in this study that Ontario courses are not only designed to be passed, but the average student should be getting a Level 3. Therefore, most students should at least obtain the credit. Schunk (1996) pointed out that classroom tasks are designed for students to be successful. However, widespread grade inflation could be harmful in the long run (Zirkel, 1999; Howley, Kusimo and Parrot, 2001). To be clear, all the participants in the study had to assign a failing grade at some point in their careers. The Earl, et al., (2010) study found that only 11% of participants felt it was not a priority for all students to pass, although the participants recognized the importance is maximizing pass rates. In the present study, finding ways to justify a passing grade was a significant part of Heuristic Assessment.

4.7.6 The Reliability and Validity of Professional Judgement Summary

Professional judgement will always need to address the subjective nature of assessment. Ensuring construct validity in classroom assessment can produce dependable overall grades. The meaning of grades is situated, and therefore we should look at the plausibility of grades and the message that should be inferred from a student's score (Allal, 2013). Validity in educational measurement is dynamic; an evaluation of a learning process is best understood in comparison to the situation surrounding what is being measured (Moss, 2003). The aim is to have a sufficient amount of information to complete the objective of determining a

valid overall score (Smith, 2003). Professional judgement is used to help decide how much evidence is required to establish sufficiency. Such ideas are outside of the realm of the traditional understanding of reliability and validity, so this paradigm is misapplied to the situation of classroom assessment. Trying to force a conventional, scientific-measurement approach to evaluating student learning does not work. Instead, matters such as social consequences, how the results will be interpreted and used, enter into the subjective equation. Furthermore, teachers will try to simplify a complex process, or what this study termed Heuristic Assessment or what the literature calls construct irrelevance. In the interest of being fair and equitable, teachers will adjust an overall grade a few percentage points, usually to increase grades or ensure a student will pass. Unfortunately, these decisions are often based on non-achievement factors. Although the grade is meant to be transparent, efforts are made to shadow exactly how the grade is decided upon. Educators have the challenging profession of managing a classroom full of students, following directions from the administration, and negotiating with stakeholders. Furthermore, teachers need to figure out what to do about missing assignments, gathering different forms of evidence for different purposes, and establishing a student's most consistent level of achievement on a 100-point scale. Informed professional judgement guides the educator in all arenas of the situation. However, when it comes to determining the report card grade, professional judgement should not be an after-the-fact last step, but incorporated in the evaluation process itself. Therefore, Ontario needs to ameliorate assessment and evaluation practices of its teachers by better clarifying the polysomic term professional judgement.

4.8 Literature Review and Discussion Summary

Whether the subject literature was discussing American High School teachers (e.g., Stiggins, Frisbie and Griswold, 1989), Swiss Elementary Teachers (Allal, 1988; 2013), or Ontario Secondary School teachers (e.g., Earl, et al., 2010), the information was often reminiscent of what the participants reported in this study. In the situated classroom, there are many challenges to recording accurate, or plausible, achievement. Many factors feel like they matter in the interest of construct validity, only to expose construct irrelevance under closer examination. Furthermore, they often fail to live up to standards of transparency. Solutions such as more teacher moderation are promising, but again, there are the challenges of the real world; a clearer case of how summative assessment is to improve student learning needs to be made.

Table 4.1: Literature Review and Discussion Sources provides an overview the sources cited in this chapter. The table simply provides information on the author(s), title of the text, and a note on the methodology and/or methods approach. As previously stated, much of the literature expands on previous findings in the form of reviews, discussions, and position papers. Furthermore, authors have often taken a quantitative approach to collecting data. The present study showed how a constructive, qualitative approach can be used to gather data and make a contribution to the field of knowledge. At the same time, it is important to be aware of the field of knowledge by examining the extant literature, and identifying where the present study fits into this field. In short, this study has expanded on research into professional judgement, and how difficult it is to remain

objective on sensitive issues such as evaluation – especially when the classroom

teacher has power over the situation.

Table 4.1: Literature Review and Discussion Sources

Author(s), Year	Title	Methodology / Methods
Aaronsohn, Holmes, Foley and Wallowitz, 1994	*"Teacher-pleasing", traditional grading – and learning? A collaborative qualitative study*	Interviews with American elementary, middle, high school, and university students
Airasian, 1991	Perspectives on measurement instruction	Position paper
Airasian, 1994	*Classroom assessment*, 2nd end	Textbook
Airasian, 1996	*Assessment in the classroom*	Textbook
Airasian, Engemann, and Gallagher, 2007	*Classroom assessment: concepts and applications*	Textbook
Airasian and Jones, 1993	The teacher as applied measurer: realities of classroom measurement and assessment	Position Paper
Allal, 1988	Quantitative and qualitative components of teachers' evaluation strategies	Interviews with 45 Geneva Elementary teachers from three schools
Allal, 2013	Teachers' professional judgement in assessment: a cognitive act and a socially situated practice	Interviews with ten Grade 6 teachers from ten Swiss schools
Anders and Richardson, 1992	Teacher as game-show host, bookkeeper, or judge? Challenges, contradictions, and consequences of accountability	Position paper
Artus, 2014	EQAO scores have value: Education Minister Liz Sandals	Newspaper article
Baker and O'Neil, 1994	Performance assessment and equity: a view from the USA	Position paper
Barnes, 1985	A study of classroom pupil evaluation: the missing link in teacher education	Interviews with 20 Teachers' College instructors and 20 student teachers in the southwestern United States
Black, 1993	Formative and summative assessment by teachers	Literature review and discussion
Black, 2003	Testing, testing: listening to the past and looking to the future	Literature review and discussion
Black, 2015	Formative assessment – an optimistic but incomplete vision	Literature review and discussion
Black and Wiliam, 2003	In praise of educational research: formative assessment	Literature review and discussion
Black and Wiliam, 2012	The reliability of assessments	Review based on four major studies
Bond, 1995	Unintended consequences of performance assessment: issues of bias and fairness	Literature review and discussion
Brantlinger, 1993	*The politics of social class in secondary school: views of affluent and impoverished youth*	Textbook
Brookhart, 1991	Grading practices and validity	Letter to *Educational Measurement: Issues and Practice*
Brookhart, 1993	Teachers' grading practices: meaning and values	Mixed method analysis of 84 American teachers, 40 with and 44 without assessment measurement training
Brookhart, 1994	Teachers' grading: practice and theory	Mixed method analysis of 84 American teachers, 40 with and 44 without assessment measurement training
Brookhart, 1997	A theoretical framework for the role of classroom assessment in motivating student effort and achievement	Literature review and discussion
Brookhart, 1999	Teaching about communicating assessment results and grading	Position paper
Brookhart, 2001	Successful students' formative and summative assessment use of assessment information	Interviewed fifty American English and Anatomy High School students
Brookhart, 2004	*Grading*	Textbook
Brookhart, 2011	Educational assessment knowledge and skills for teachers	Position paper

Brookhart and Freeman, 1992	Characteristics of entering teacher candidates	Literature review and discussion
Canady and Hotchkiss, 1989	It's a good score: just a bad grade	Position paper
Carifio and Carey, 2009	A critical examination of current minimum grading policy recommendations	Literature review and discussion
Cherryholmes, 1988	Construct validity and discourses of research	Literature review and discussion
Cizek, Fitzgerald and Rachor, 1995	Teachers' assessment practices: preparation, isolation and the kitchen sink	Surveyed 60 American middle and high school teachers
Cooksey, Freebody and Wyatt-Smith, 2007	Assessment as judgement-in-context: analysing how teachers evaluate students' writing	Analysis of Australian teachers' assessment judgement
Cooper, 2007	*Talk about assessment: strategies and tools to improve learning*	Textbook
Cooper, 2010	*Researching assessment and evaluation*	Personal correspondence
Cooper, 2011	*Redefining fair: how to plan, assess, and grade for excellence in mixed-ability*	Textbook
Cooper, O'Connor, and Wakeman, 2009	Redefining fair: assessment and grading in the 21st century	Position paper
Costello and McKillop, 2000	Dealing with late and absences	Position paper
Cronbach, 1989	Construct validation after thirty years	Literature review and discussion
Crooks, 1988	The impact of classroom evaluation practices on students	Literature review and discussion
Crooks, Kane, and Cohen, 1996	Threats to the valid use of assessments	Position paper
Cross and Frary, 1999	Hodgepodge grading: endorsed by students and teachers alike	Surveyed 397 middle and high school teachers and 8664 students
Cummings, et al., 2006	*Teacher judgement: building an evidentiary base for quality literacy and numeracy education*	Textbook
Datnow and Hubbard, 2015	Teachers' use of assessment data to inform instruction: lessons from the past and prospects for the future	Literature review and discussion
Daugherty, et al., 2012	Alternative perspectives on learning outcomes: challenges for assessment	Position paper
Davies, 2000	*Making classroom assessment work*	Textbook
Davies, 2007	*Making classroom assessment work.* 2nd ed	Textbook
Davies, 2011	*Making classroom assessment work.* 3rd ed	Textbook
Dress man, Journell and Mann, 2012	Teacher education: qualitative research approaches	Literature review and discussion
Duncan and Noonan, 2007	Factors affecting teachers' grading and assessment practices	Survey of 513 western Canadian Secondary School teachers
Earl, et al., 2010	*Classroom assessment in Ontario secondary schools: in teachers' hands*	Survey of 5905 Ontario teachers from all 72 boards, follow-up interviews with 72 teachers and nine principals from nine schools across the province
Ebel and Frisbie, 1991	*Essentials of educational measurement*	Textbook
Elwood and Klenowski, 2002	Creating communities of shared practice: assessment use in learning and teaching	Action research and case study
Elwood and Murphy, 2015	Assessment systems and cultural scripts: a sociocultural theoretical lens on assessment practice and products	Literature Review and Discussion
Fleer, 2015	Developing an assessment pedagogy: the tensions and struggles in re-theorizing assessment from a cultural-historical perspective	Study of eleven Australian primary teachers
Frary, Cross and Weber, 1993	Testing and grading practices and opinions of secondary teachers of academic subjects: implications for instruction in measurement	Survey of 536 Virginian High School teachers
Freebody and Wyatt-Smith, 2004	The assessment of literacy: working the zone between 'system' and 'site' validity	Position paper
Friedman and Frisbie, 1995	The influence of report cards on the validity of grades reported to parents	Wisconsin report card analysis, including 39 from kindergarten, 59 elementary, 48 middle school, and 70 high school

Friedman and Manley, 1992	Improving high school grading practices: "experts" vs. practitioners	Survey of 227 teachers, students, administration and counselors, from 35 different Wisconsin High Schools
Friedman and Troug, 1998	Evaluation of high school teachers' written grading policies	Analysis of Midwestern American teachers' grading practices
Frisbie and Waltman, 1992	Developing a personal grading plan	Module
Garcia, 1995	Equity challenges in authentically assessing students from diverse backgrounds	Position paper
Gardner, 2012a	Quality assessment practice	Position paper
Gardner, 2012b	*Assessment and learning*	Textbook
Gardner, et al., 2010	*Developing teacher assessment*	Textbook
Gill, 2013b	*Secondary evaluation policy*	Personal correspondence
Gipps, 1994	*Beyond testing: towards a theory of educational assessment*	Textbook
Gipps, 1995	What do we mean by equity in relation to assessment?	Position paper
Goldstein, 2015	Validity, science and educational measurement	Literature Review and Discussion
Good and Brophy, 2000	*Looking in classrooms*	Textbook
Graue, 1993	Integrating theory and practice through instructional assessment	Position paper
Gronlund, 1998	*Assessment of student achievement*	Textbook
Gronlund and Linn, 1990	*Measurement and evaluation in teaching*	Textbook
Gullickson, 1993	Matching measurement instruction to classroom-based evaluation: Perceived discrepancies, needs, and challenges	Literature review and discussion
Guskey, 1994	Making the grade: what benefits students?	Position paper
Guskey, 2004	"Zero" alternatives	Position paper
Guskey, 2006	It wasn't fair!" Educators' recollections of their experiences as students with grading	Position paper
Guskey and Bailey, 2010	*Developing standards-based report cards*	Textbook
Hammond, 1996	*Human judgment and social policy: irreducible uncertainty, inevitable error, unavailable injustice*	Textbook
Hargreaves, 2004	Culture, contracts and change	Position paper
Harlen, 1994	Developing public understanding of education – a role for researchers	Literature review and discussion
Harlen, 2005a	Teachers' summative practices and assessment for learning – tensions and synergies	Literature review and discussion
Harlen, 2005b	Trusting teacher's judgment: research evidence of the reliability and validity of teacher's assessment used for summative purposes	Literature review and discussion
Harlen, 2006	On the relationship between assessment for formative and summative purposes	Literature review and discussion
Harlen, 2012a	On the relationship between assessment for formative and summative purposes	Literature review and discussion
Harlen, 2012b	The role of assessment in developing motivation for learning	Literature review and discussion
Harlen and Deakin Crick, 2003	Testing and motivation for learning	Literature review and discussion
Harlen and James, 1997	Assessment and learning: differences and relationships between formative and summative assessment	Literature review and discussion
Harris and Brown, 2009	The complexity of teachers' conceptions of assessment: tensions between the needs of schools and students	Surveyed 161 New Zealand teachers (primary, intermediate, and high school) with 26 participants consenting to follow-up interviews; phenomenogical analysis
Hawe, 2003	'It's pretty difficult to fail': the reluctance of lecturers to award a failing grade	Straussian-influenced grounded theory project
Hayward, 2015	Assessment is learning: the preposition vanishes	Literature review and discussion of Scottish assessment
Herbst and Davies, 2014	*A fresh look at grading and reporting*	Textbook

Hoge and Coladarci, 1989	Teacher-based judgments of academic achievement: a review of the literature	Literature review and discussion
Howley, Kusimo and Parrot, 2001	Grading and the ethos of effort	Interviewed 52 female American Seventh Graders and their teachers
James, 1998	*Using assessment for school improvement*	Textbook
Joint Advisory Committee, 1993	*Principles for fair student assessment practices for education in Canada*	Position paper
Kane, 1982	A sampling model for validity	Literature review and discussion
Klapp, 2015	Does grading affect educational achievement? A longitudinal study	Longitudinal study of 8558 Swedish students born in 1967 and their final evaluations
Klenowski and Wyatt-Smith, 2013	*Assessment for education: standards, judgement and moderation*	Textbook
Kohn, 1993	*Punished by rewards*	Textbook
Kohn, 1996	*Beyond discipline: from compliance to community*	Textbook
Koretz, 2008	*Measuring up: what educational testing really tells us*	Textbook
Larson, 2009	Stressful, hectic, daunting: a critical policy study of the Ontario teacher performance appraisal system	Surveyed and interviewed 125 Ontario teachers (elementary, intermediate, and senior)
Linn, 1994	Performance assessment: policy promises and technical measurements standards	Literature review and discussion
Linn and Gronlund, 1995	*Measurement and assessment in teaching*	Textbook
MacLeans.ca, 2011	McGuinty boats rising graduation rates	Magazine article
Marzano, 2000	*Transforming classroom grading*	Textbook
Matanin and Tannehill, 1994	Assessment and grading in physical education	Literature review and discussion
Matteucci, et al., 2008	Teacher judgments and pupils' casual explanations: social valorization and effort-based explanations in school context	Survey of 126 Italian primary school teachers
McMahon and Jones, 2015	A comparative judgement approach to teacher assessment	Study of Irish Chemistry teachers shifting to the use of internal classroom assessment
McMillan, 2001	Secondary teachers' classroom assessment and grading practices	Surveyed 1483 Grade 6-12 teachers within 53 Virginian schools
McMillan, 2003	Understanding and improving teachers' classroom assessment decision making: implications for theory and practice	Literature review and discussion
McMillan, 2008	*Assessment essentials for standards-based education*, 2nd ed	Textbook
McMillan, Hellsten, and Klinger, 2011	*Classroom assessment: principles and practices for effective standards-based instruction*	Textbook
McMillan, Myran, and Workman, 2002	Elementary teachers' classroom assessment and grading practices	Survey of 900 American Elementary teachers
McMillan and Nash, 2000	Teacher classroom assessment and grading decision making	Interviews with 27 High School teachers, and one elementary, from twelve Virginian schools
Mehan, 1998	The study of social interaction in educational settings: accomplishments and unresolved issues	Literature review and discussion
Messick, 1989a	Validity	Position paper
Messick, 1989b	Meanings and values in test validation: the science and ethics of assessment	Literature review and discussion
Messick, 1995	Standards of validity and validity of standards in performance assessment	Position paper
Miller and Frederick, 1999	How does grounded theory explain?	Literature review and discussion
Miller and Linn, 2000	Validation of performance-based assessments	Literature review and discussion
Mislevy, 1993	Foundations of a new test theory	Position paper
Morrison and Wylie, 1999	Why national curriculum testing is founded on a methodological thought disorder	Position paper
Moss, 1992	Shifting conceptions of validity in educational assessment: implications for performance assessment	Literature review and discussion
Moss, 1994	Can there be validity without reliability?	Literature review and discussion
Moss, 1995	Themes and variations in validity theory	Literature review and discussion

Moss, 2003	Conceptualizing validity for classroom assessment	Position paper
Moss, 2012	Exploring the macro-micro dynamic in data use practice	Position paper
Musial, et al., 2009	*Foundations of meaningful educational assessment*	Textbook
Newmann, 1990	Higher order thinking in teaching and social studies: a rationale for the assessment of classroom thoughtfulness	Literature review and discussion
Newton, 2005	The public understanding of measurement inaccuracy	Literature review and discussion
Newton, 2007	Clarifying the purposes of educational assessment	Literature review and discussion
Newton, 2012	Validity, purpose and the recycling of results from educational assessments	Literature review and discussion
Newton and Meadows, 2011	Marking quality within test and examination systems	Editorial
Nitko, 2001	*Educational assessment of students*	Textbook
O'Connor, 2000	Grading: an exercise in professional judgment	Position paper
O'Connor, 2002	*How to grade for learning: linking grades to standards*	Textbook
O'Connor, 2007	*A repair kit for grading: 15 fixes for broken grades*	Textbook
O'Connor, 2009	*How to grade for learning: K-12*. 3rd ed	Textbook
O'Connor, 2010	*Researching assessment and evaluation*	Personal correspondence
O'Connor and Wormeli, 2011	Reporting student learning	Position paper
O'Donovan, Price and Rust, 2004	Know what I mean? Enhancing student understanding of assessment standards and criteria	Position paper
Office, 2015	More Ontario students graduating high school than ever before	Statement from the Premier
Ontario: Ministry of Education, 2010c	*Student success / learning to 18*	Ministry statement
Ontario: Ministry of Education, 2013a	*Getting results: Ontario's graduation rates*	Ministry statement
Plake and Impara, 1993	Assessment competencies of teachers: a national survey	Survey of seventy American teachers
Plake and Impara, 1997	Teacher assessment literacy: what do teachers know about assessment?	Literature review and discussion
Pilkington, 2004	*Practice and the human sciences*	Textbook
Pollio and Humphreys, 1988	Grading students	Literature review and discussion
Pollitt, 2012	The method of adaptive comparative judgement	Literature review and discussion
Popham, 2003	Seeking redemption for our psychometric sins	Commentary
Popham, 2009	Assessment literacy for teachers: faddish or fundamental?	Position paper
Pugsley, 2012	Walk this way, talk this way: qualitative research on professional education	Literature review and discussion
Reeves, 2004	The case against zero	Position paper
Reeves, 2010	*Elements of grading*	Textbook
Reiman and Johnson, 2003	Promoting teacher professional judgment	Literature review and discussion
Ross and Kostuch, 2011	Consistency of report card grades and external assessments in a Canadian province	Analysis of the external test scores of 15,942 Ontario Elementary students and took of sample of 480 student report cards for comparison
Rowe and Hill, 1996	Assessing, recording and reporting students' educational progress: the case for 'subject profiles'	Analysis of Australian student profiles as an efficient form of record keeping
Rushowy, 2015	High school grad rates raise slightly in Ontario	Newspaper article
Sadler, 1985	The origins and functions of evaluative criteria	Literature review and discussion

Sadler, 1987	Specifying and promulgating achievement standards	Literature review and discussion
Sadler, 1998	Formative assessment and the design of instructional systems	Literature review and discussion
Sadler, 2009	Indeterminacy in the use of preset criteria for assessment and grading	Literature review and discussion
Scates, 1943	Difference between measurement criteria of pure scientists and of classroom teachers	Mixed method analysis of classroom assessment
Schunk, 1996	Goal and self-evaluative influences during children's cognitive skill learning	Multiple case study of American Grade Four students
Schwandt, 1996	Farewell to criteriology	Position paper
Scott et al., 2014	Fair and equitable assessment practices for all students	Surveyed 3312 Albertan stakeholders as a follow up to Joint (1993)
Selby and Murphy, 1992	Graded or degraded: perceptions of letter-grading for mainstreamed learning-disabled students	Study of six Canadian elementary students
Seligman, 1998	*Learned optimism*	Textbook
Shepard, 1993	Evaluating test validity	Literature review and discussion
Shepard, 1995	Using assessment to improve learning	Study of fourteen American elementary math teachers
Shepard, 1997	*Measuring achievement: what does it mean to test for robust understandings?*	Textbook
Shepard, 2000a	The role of assessment in a learning culture	Literature review and discussion
Shepard, 2000b	The role of classroom assessment in teaching and learning	Literature review and discussion
Shepard, 2009	Commentary: Evaluating the validity of formative and interim assessment	Position Paper
Simon, et al., 2010	A secondary school teacher's description of the process of determining report card grades	Case study of one Grade 10 Ontario French Math teacher
Smith, 2003	Reconsidering reliability for classroom assessment and grading	Literature review and discussion
Smith, 2014	Why principals may be the key to your child's success at school	Newspaper article
Speck, 1998	Unveiling some of the mystery of professional judgment in classroom assessment	Literature review and discussion
Stake, 2004	*Standards-based responsive evaluation*	Textbook
Steffenhagen, 2010	Zeros now possible when Ontario students missed deadlines	Newspaper article
Stiggins, 1988	Revitalizing classroom assessment: the highest instructional priority	Position paper
Stiggins, 1992	High quality classroom assessment: what does it really mean?	Position paper
Stiggins, 1997	*Student-centered classroom assessment*	Textbook
Stiggins, 1999	Evaluating classroom assessment training in teacher education programs	Position paper
Stiggins, 2001a	*Student-involved classroom assessment.* 3rd ed	Textbook
Stiggins, 2001b	The unfulfilled promise of classroom assessment	Literature review and discussion
Stiggins, 2005	*Student-involved assessment for learning*	Textbook
Stiggins and Bridgeford, 1985	The ecology of classroom assessment	Survey of 228 American teachers from across the country representing all grade levels
Stiggins and Conklin, 1992	*In teachers' hands: investigating the practices of classroom assessment*	Textbook
Stiggins and Duke, 1991	District grading policies and their potential impact on at-risk students	Conference paper
Stiggins, Frisbie and Griswold, 1989	Inside high school grading practices: building a research agenda	Case study of fifteen American High School teachers
Stobart, 2005	Fairness in multicultural student assessment systems	Literature review and discussion
Stobart, 2008	*Testing times: the uses and abuses of assessment*	Textbook
Sutton, 1991	*Assessment: a framework for teachers*	Textbook

Terwilliger, 1989	Classroom standard setting and grading practices	Position paper
Thornberg, 2007	Inconsistencies in everyday patterns of school rules	Ethnographic study of two Swedish elementary schools
Thornberg, 2009	The moral construction of the good pupil embedded in school rules	Case study of two Swedish elementary schools
Tierney, 2012	Why do so many teachers quit their jobs? Because they hate their bosses	Magazine article
Tittle, 1989	Validity: whose construction is it in the testing and learning context	Literature review and discussion
Troug and Friedman, 1996	Evaluating high school teachers' written grading polices	Focus group
Trouilloud, et al., 2002	The influence of teacher expectations on student achievement in physical education classes: Pygmalion revisited	Multiple serves of 173 French students and seven teachers
Western and Northern Protocol for Collaboration in Education, 2006	*Rethinking classroom assessment with purpose in mind*	Position paper
Volante, 2006	Reducing bias in classroom assessment and evaluation	Position paper
Walvoord and Anderson, 1998	*Effective grading: a tool for learning and assessment*	Textbook
Wenger, 1998	*Communities of practice: learning, meaning, and identity*	Textbook
Whittington, 1999	Making room for values and fairness: teaching reliability and validity in the classroom context	Position paper
Wiggins, 1993	Assessment, authenticity, context, and validity	Position paper
Wilen, et al., 2004	*Dynamics of effective teaching*	Textbook
Wilson, 1998	Educational standards and the problem of error	Literature review and discussion
Wormeli, 2006b	Turning zeros to 60s	Position paper
Wyatt-Smith, 1999	Reading for assessment: how teachers ascribe meaning and value to student writing	Interviews with Queensland, Australia teachers on their grading practices
Wyatt-Smith and Castleton, 2005	Examining how teachers judge student writing: an Australian case study	Case study
Wyatt-Smith and Klenowski, 2013	Explicit, latent and meta-criteria: types of criteria at play in professional judgement practice	Interviews with 89 Elementary and Intermediate Australian teachers from 49 schools
Wyatt-Smith, Klenowski and Gunn, 2010	The centrality of teachers' judgement practices in assessment: a study of standards in moderation	Mixed methods of 15 Australian (Queensland) teachers
Zeidner, 1992	Key facets of classroom grading: A comparison of teacher and student perspectives	Survey of 402 northern Israeli students and 174 teachers
Zirkel, 1999	Grade inflation: a leadership opportunity for schools of education?	Position paper
Zwaagstra, 2012a	Keep no-zero policies out of schools in Nova Scotia	Newspaper article
Zwaagstra, 2012b	The folly of 'no-zero' policies in schools	Newspaper article
Zhu and Urhahne, 2015	Teacher judgements of students' foreign-language achievement	Study of sixteen Chinese English teachers of 555 students

This study has spoken on the subject of interpretation and accuracy. Part of the criticism of a constructivist approach is the assertion that all interpretations are plausible as knowledge is co-created. Furthermore, some mistake this claim

that a researcher can do whatever s/he likes and claim validation. The

Methodology and Methods chapter went to lengths to demonstrate that such

statements are misleading and dismissive of the potential contribution to

knowledge of the constructivist approach. A researcher still needs to explain

carefully how conclusions were reached and how criteria were constructed in order

to show the trustworthiness of a conclusion. Likewise, an educator in an internal

assessment system such as Ontario cannot determine a grade on professional

judgement alone. The teacher is responsible to many stakeholders to ensure a

grade is imprimatur. However, as we read in the analysis, the participants had to

address many real world issues, in addition to achievement, during the evaluation

process. Consequently, professional judgement became Heuristic Assessment to

help pinpoint the report card grade. An examination of the literature showed that

the experience of the participants is not unusual. Furthermore, we can use the

research of others to formulate a plan on how to improve upon the use of

professional judgement as a necessary part of assessment and evaluation.

CHAPTER FIVE: THE FUTURE OF INFORMED PROFESSIONAL JUDGEMENT IN ONTARIO

5.1 Introduction

This thesis has made a contribution to the field of knowledge by using constructivist grounded theory to investigate the under-researched arena of the use of professional judgement and final report cards in the social world of Ontario's Secondary Schools. The concept of professional judgement is advertised as the lynch pin in provincial assessment, yet the term itself is not truly understood. Consequently, when negotiating a student's achievement in the form of a symbolic percentage marker, the participants reverted to what was termed Heuristic Assessment: a personal approach to grading based on an understanding of policy, taking the local situation into account, and reflecting on both achievement and behaviour. Heuristic Assessment is further encouraged by local administrations who desire positive report card results. The literature review and discussion framed the findings as consistent with extant ideas while expanding upon the continuing problematic situation of classroom assessment. In this final chapter, we will review how the situation of Heuristic Assessment can be addressed to establish a more consistent approach to informed professional judgement to improve the quality of final report card evaluation practices.

Miles, Huberman, and Saldaňa (2014, p.66) advised that a study's "write-up should also clearly specify how you envision your findings being used by others, especially for policy and practice." This final chapter provides an overview of how awareness gained from the research could benefit assessment practices in Ontario. The participants in the study demonstrated many good

qualities that, when combined, provide an exemplar for other educators in the
province to follow. This exemplar could help to inform teaching training and day-
to-day practices. Furthermore, *Growing Success* encompasses decades of positive
assessment reform that can serve as a model for other educational systems.
Indeed, establishing a unified assessment policy is a feat in itself. At the same
time, vagueness in the policy must be corrected with a series of pithy statements
that do not attempt to change the existing policy nor take away from board
independence – but take the next step forward towards consistency of assessment
practices that will add reliability and validity to final report card grades. Great
work has been done on how to use assessment and evaluation to improve student
learning, but how to take evidence of learning and determine a report card grade
using informed professional judgement must be clarified. A series of
recommendations are made that could appear in a second edition of *Growing
Success*. Additional assessment literature has been provided to help support these
recommendations. The chapter concludes with addressing weaknesses in the
study, and how they could be addressed in follow-up studies.

5.2 The Ideal Approach to Informed Professional Judgement

The participants displayed many assessment merits. Although there were
several concerns with particular steps taken, such as those that could lead to
hodgepodge grading or other evaluation epiphenomena, many of these procedures
were the result of shadowed practices long-embedded in school cultures. When
the actions of the participants are brought together as well, the voices presented in
Chapter Three illustrated an ideal approach to evaluation as informed professional

judgement can be modeled. For instance, Catherine and Jordan did not believe

that punitive measures should be mandatory, but should be left to the discretion of

the classroom teacher. Similarly, Dirk felt that an educator should show empathy

before reverting to such measures. As James believed, it is important to work with

students to maximize their potential. This approach includes emphasizing the

importance of completing assignments instead of automatically reverting to a

punishment. Moreover, any punitive measures must be tracked to ensure that,

when the overall mark is determined, other evidence of achievement is considered.

In regards to students who are resistant to the learning process, Miguel and Sally

said it is essential to think holistically. Smitty pointed out that educators must

look beyond negative behaviours to see a student's potential. As Helen said, an

educator must be open-minded. There are countless factors that can be the true

cause of negative behaviour. It is important not to let negative behaviour distract

from evidence of achievement. Oliver observed some students are naturally shy

and may be hesitant to participate. Therefore, educators must collect evidence of

learning in multiple ways. That said, Oscar felt that a student also has a

responsibility to reach out for help when needed. The teacher must not give the

impression that they are not available for help when requested. In other words,

both the student and teacher, as the lead learner, share in a commitment to success.

The above only begins to summarize the challenges to classroom

evaluation and the challenges to professional judgement. Eileen showed the

importance of having confidence in one's teaching and evaluation skills. Lucy

also showed having professional self-reliance includes not needing a computer to

determine a student's most consistent achievement with special consideration for more recent. This level of achievement should be obvious to the teacher who spent an entire course interacting with a student. At the same time, records need to be kept as memory alone is not reliable and lacks transparency. Both Winnie and Larry explained that determining a level for achievement involves a reflection of all the evidence, and to allow the best evidence of learning to surface. On another important note, Lorrie demonstrated the power of teacher moderation and having a coterie of professionals to help determine a student's grade. This idea extends to Corey, Derek, and Jerry explaining how an educator needs to be able to provide to students and other stakeholders transparency on how a final grade was determined. As a new educator, Denise immediately saw the importance of recognizing the classroom dynamic because every group of students will be different. Or as Murray said, an educator needs to take in the whole situation in order to be fair and equitable.

Although the administration has an advisory and overseeing role, as Greg discussed, an educator needs to feel trust and support. After all, Lisa believed, professional judgement is an expression of credentials. At the same time, Harry explained, in the assessment process educators need to work in partnership with the administration. George, reflecting on his experience as an educator, discussed the difficulty of assessing subjective subjects like social science, but asserted that teachers must use their skills. Guidance from the administration is an important part of this process.

Applying informed professional judgement is the ability to work with the range of variance. Students will display various levels of achievement, combined with various behaviours. Furthermore, there will be Ministry directives and procedures recommended by the local administration. There will be challenges to what constitutes good evaluation practices, but there is also support from colleagues. To be informed is to take all of these factors into account, while professional judgement determines what is needed to determine a grade, and what information serves other purposes ranging from discipline to additional student support. Often what feels like the right action to take with a particular student interferes with overall reliability and validity. Again, it is not that certain evidence is ignored entirely, since it may be a sign of a problem that needs to be addressed. What needs to be stressed is there are elements that belong in a grade, and other factors that are treated in a different manner.

5.3 Improving the Quality of Informed Professional Judgement

The OME's definition of professional judgement suggested that it is something that should improve over time with teaching experience. However, if professional judgement only relied on experience, then a new teacher such as Denise would not be able to exercise this vital component. Therefore, by better understanding the concept of professional judgement, we can explore now it can be explained to both new and experienced teachers to improve their practices. The goal of what Stiggins (2001a) called "high-quality classroom assessment" could be achieved by reflecting on construct validity: assessment and evaluation practices should incorporate the connotations of reliability and validity by using the content

standards to drive instruction and collect a sufficient amount of evidence; it would

also include the social consequences of grading, including how the public would

react to a more honest approach to the imperfections of assessment and what

evaluation is truly meant to do. The evidence presented in both the analysis and

literature review demonstrated that it would be naïve to suggest there is a unified

approach to assessment and evaluation that could solve all of its problems. That

said, a thorough understanding of the implications of professional judgement could

do much to address many issues presented in this report. I believe that by openly

discussing professional judgement and these connecting issues, both new and

veteran educators could greatly enhance their assessment skills.

Although there are currently no plans for a second edition of *Growing*

Success, the OME has released follow-up documents that have further emphasized

the need for informed professional judgement (Ontario, 2011; 2013b). These

documents elaborate on the greater framework of education in the province, and

the essential role of different forms of assessment working together to improve

student learning. However, when it comes to what informed professional

judgement is as a concept, references are made back to definitions provided by

Growing Success. Consequently, how an educator is to take evidence of

achievement and quantify it as an overall grade has gone unexplored. Granted,

how classroom learning benefits students is of paramount concern. But as long as

grades still maintain the current level of cultural currency, social value, acceptance

to post-secondary institutions, then educators are obligated to have a firm

understanding of how to ensure authenticity in the grades they assign.

Furthermore, the move from a one-year to a two-year teaching college program in Ontario has no plans to formally address issues of assessment and evaluation. It is imperative that the province instructs Teacher Colleges to enhance discussion of how to determine report card grades. Moreover, administrations must ensure staff understands the fundamentals of assessment and evaluation.

5.3.1 Proposed Amendments to Evaluation Training in Ontario

The data demonstrated that Ontario educators tend to enter the professional without sufficient assessment literacy and learn to do the report card evaluations independently (see Section 3.2.1). Dressman, Journell and Mann (2012) explained that the literature on teacher training is a relatively new and developing field. Therefore, providing specific advice on how best to improve evaluation training requires further refinement. Understanding the profession makes more sense with actual job experience (McMillan, 2003). Nevertheless, the topic needs greater attention in teacher training. Despite the frequency of summative assessments in the Secondary School classroom to cover the overall expectations and 70% of the course grade, not to mention the massive amount of man-hours proper evaluation demands, researchers have found parsimonious instruction on proper assessment practices in teacher training programs and in-service professional development (Schafer and Lissitz, 1987; Crooks, 1988; Stiggins and Conklin, 1992; Whittington, 1999; Stiggins, 2001b; Volante, 2007; Popham, 2009). Brown (2004) and Volante (2006) pointed out that educator training can be otiose on covering proper assessment practices, and that Teacher Colleges have a

responsibility to help educators with the development of their professional judgement.

Even though professional judgement benefits from experience, there are points that a new educator can incorporate into early practices. For instance, Sadler (1998) pointed out there are the three elements to teacher judgement in regards to evaluation: looking at the learner, knowing the standards, and determining the grade. Such assessment literacy is needed as part of informed professional judgement (Crooks, 1988; Hoge and Coladari, 1989; Stiggins, 1988; 1991; 1999; Stiggins and Conklin, 1992; McMillan, 2001; 2008; Popham, 2004; Gardner, et al., 2010). Although OME documents have stressed the importance of understanding the learning and using the content standards to drive assessment, how actually to determine a grade remains vague. Such instruction must become a visible part of teacher education. For example, more scenarios need to be presented to teacher candidates including data of fictional students' most consistent and more recent achievement, along with personal details about the example student, in order to illustrate points made in *Growing Success*. Teacher candidates could then be invited to discuss how they would go about determining the report card grade. Although these discussions would not impact a student the candidates have taught, starting a dialogue would at least introduce future teachers to the implications of their grading decisions such as social consequences.

5.3.2 Teacher Professional Development

The task of evaluation instruction cannot be left to Teachers' Colleges alone. It needs to be an active part of school culture. Experienced educators could be mandated to help new colleagues develop understanding of what good assessment practices look like, how to apply them to the curriculum, monitor for bias such as the halo effect, not using assessment as a means to control behaviour, and communicating achievement (Frisbie and Waltman, 1992; Brookhart, 1999; Ingersoll, 1999; Brown, Glasswell and Harland, 2004). Granted, the NTIP is meant to provide such instruction but comments made by the participants suggested it is not effective. Administrators would also need to ensure that these coaches know how to eliminate miscalculations in assessment and explain the nature of measurement error (Stiggins, 2001b; Newton, 2005). Otherwise, educators are likely to evaluate based on misconceptions (Guskey, 1996). Teacher professional development is a career-long project and proper assessment practices is a topic, in some form or fashion, that needs to be reviewed each school year.

However, educators have been known to be resistant to changing views on assessment and evaluation, or even discussing good practices with colleagues (Bond, 1995). Furthermore, even when teachers are aware of measurement error, they are likely to favour the student to be on the safe side (Newton, 2005). Brookhart (1994) looked at nineteen previous studies and found, even when adequate training had been received, there were drastic differences in how educators applied assessment knowledge. It is not essential that all educators be voluble about assessment, but a *verstehen* of how evaluation works and should be

practiced is needed. Administrations need to take responsibility to improve the professional development of all teachers in regards to assessment and evaluation, and have a means to measure the steps they have taken to ensure good practices.

5.3.3 Promoting Informed Professional Judgement in School Culture

A recurring topic in this study was the importance of school culture, and how the administration can influence the assessment situation. The importance of the direction provided by a well-run administration has been noted in the literature (Beach, 1992; Joint Committee, 1993; Howley, Kusimo and Parrot, 2001; Costa and Kallick, 2001; Harlen and Deakin Crick, 2003; Western, 2006; Earl, et al., 2010; Gardner, et al., 2010). Local administrations need to work with initiatives coming down from the board, but we have seen evidence that schools have room to maneuver when it comes to interpreting policy. However, individual schools could have a better application and explanation of proper assessment practices. The creation of an awareness of good, shared practices would lead to a greater consistency in the province and a more efficient system of evaluation.

Similar to this study, McMillan, Hellsten and Klinger (2011) claimed grading systems could be different within the same school. When educators are not communicating with each other, even within the same department, inconsistency in assessment practices is a likely result (Carifio and Carey, 2009; O'Connor and Wormeli, 2011). Granted, every class is different and students themselves influence the school culture; therefore, assessment procedures need to remain flexible (Stiggins and Conklin, 1992; Brookhart, 1997; Stiggins, 2001a).

However, flexibility does not mean that decisions do not correspond to greater directives within the situation that give these decisions construct validity. Howley, Kusimo and Parrot (2001) reflected on situational elements in their study and were surprised at the lack of studies considering the connection between school environment and grading practices. Teachers should be familiar with the written policy and talk to the administration about points of contention (Frisbie and Waltman, 1992). The lack of such conversations was noted the participants, as Heuristic Assessment was applied in cases on disagreement and/or uncertainty. The study does not suggest there are easy answers but emphasizes the need to establish a healthy dialogue on assessment in each and every Ontario school as an excellent start.

Granted, evaluation is a divisive topic. With so many competing views, establishing a conversation on assessment practices brings up its own challenges. Brantlinger (1993) explained that every teacher brings his or her personal background to the situation of the school. Prior experience may conflict with what the administration wants, and this conflict can lead to shadowed practices in the form of Heuristic Assessment. Again, the participants in this study were trying to find ways to both simplify the process and benefit students. Allal (2013) found that the work setting was a major influence on the development of professional judgement. Administrations need to engage in the challenging steps of explaining the basics of good assessment practices. Furthermore, if the deeper issues causing assessment construct irrelevance in the school environment are identified, then the impact on the school climate could be addressed. After all, teachers need the

guidance of a good administration (Cooper, O'Connor, and Wakeman, 2009;

Black and Wiliam, 2012; Pugsley, 2012; Tierney, 2012; Smith, 2014). New

educators, in particular, overwhelmed with the realities of the classroom, can

quickly become disconnected. Consequently, practices such as assessment could

be looked at as a job to get done instead of an ongoing process essential to

learning. The administration, as well as veteran colleagues, need to succor such

teachers. At the same time, the administration does not need to be pedantic over

the situation, but should empower teachers to use professional judgement. Davies

(2000) recognized that there must be administrative oversight. However, like our

participants, she believed that while decisions need to be consistent with

guidelines, procedures must address what is best for the individual student. What

is best does not necessarily mean what will maximize the grade, but a decision

based on construct validity, situated circumstances, and offers to improve student

learning.

5.3.4 Addressing Issues of Transparency

Throughout this study we have also seen how transparency can be

problematic. Outsiders could baulk at an explanation of evaluation as it can come

across as perfidy to those unfamiliar with proper assessment practices.

Consequently, it is difficult to initiate public debate about assessment when it can

be an emotional topic (Levin, 2004). Some fear that an attempt at full

transparency could actually cripple an education system. However, Newton

(2003; 2005) has a more sanguine view. If properly approached, he believed the

public could comprehend what assessment is and what it is meant to do. The key

is to explain how to interpret an evaluation as an approximation at a particular

point in time, and not portray this estimation as an absolute. However, due to the

social consequences of a report card grade, how to explain that it is not an absolute

poses a major obstacle. That said, we should not give up on pursuing greater

transparency in evaluation as a means to combat Heuristic Assessment (Gardner,

et al., 2010; Allal, 2013). The ongoing point made in this final chapter is that a

more open dialogue on assessment practices, between administration and staff, and

teachers and students/parents, could confront the problems that have plagued good

assessment practices. The first step towards improved informed professional

judgement is greater transparency in assessment and evaluation.

5.3.5 Use of Evaluation Software

The analysis demonstrated the strong influence evaluation software has in

Ontario schools (see Section 3.2.2). Although such software was encouraged in

quondam policy (Ontario, 1999a), there is no mention of it in *Growing Success*.

There is nothing pedagogically wrong with using evaluation software as a tool –

except when it supersedes professional judgement. Many of the participants

reported the need for final report cards to match the computer calculation.

Consequently, some participants changed summative records in order to be

compliant with this administrative guideline.

Friedman and Manley (1992) highly recommended the use of evaluation

software because it made evaluation more objective. Likewise, Friedman and

Frisbie (1995) found an unchallenged belief that software made grading more

objective. Earl, et al., 2010 found that 83% of Ontario teachers regularly use software to help determine grades. Black-Meddings, et al., (2010), in a paper for the Ontario Teachers Federation outlining *Growing Success*, spoke against computer grading as it could interfere with professional judgement. Likewise, Guskey (2002; 2013) has also warned about the use of evaluation software. Evidence from the participants suggested a general misunderstanding on how the software calculates grades. In most cases, they just entered individual scores into the software and noted what was reported. When it came time to do the final report card, respondents often adjusted scores based on Heuristic Assessment. More specifically, adjustments were based more on memory and impressions of the student. When teachers are unsure of how the score was calculated in the first place, it makes the determination of the grade lose further construct validity.

This report is not recommending the abandonment of evaluation software. Instead, it is emphasizing the need to better understand what the software is doing with the data entered. If educators are instructed input individual assignment marks, it is necessary to explain what the program is doing with the grades. Going back to Smitty's interview, Ontario teachers should not blindly follow what the computer says. Instead, they should recognize what is meant by mean, median, and mode, and use the software as a tool to help determine a student's most consistent grade with special consideration for more recent achievement. Again, when other stakeholders are brought in on this process, the result should be improved professional judgement and not the Heuristic Assessment demonstrated in this study.

5.3.6 The Use of Average

To elaborate on the previous section, Ontario teachers are to determine a student's most consistent level of achievement with special consideration for more recent achievement in order to achieve a stereoscopic view of the student's work. However, when participants are using average as a base, this criterion is ignored. Average includes all of a student's scores, thus disregarding the concept of most consistent achievement and making earlier scores immutable. Many participants report weighing assignments later in the course heavier, but this is a misinterpretation of more recent achievement as it gives automatic consideration as opposed to special consideration. O'Connor (2000; 2002; 2007; 2009; 2010) and others (Davies, 2000; Marzano, 2000; Reeves, 2010; Wormeli, 2006b) have opposed the use of average as a shibboleth. It is one of the reasons why percentage grades, and report cards in general, are misunderstood (Brookhart, 1999; 2004; 2011). Western (2006) warned that the use of average was a flawed approach to evaluation, but again, *Growing Success* did not take a stand on what actual calculation method should be used. Musial, et al., (2009) lauded the Ontario system for switching from average to most consistent / more recent and instructs American schools to do the same. However, this study shows that the use of average is a derivative of older assessment practices.

Instead of using average, median has been recommended by several noted researchers (Airasian, 1991; Marzano, 2000; O'Connor, 2002; Brookhart, 2004). Median is relatively easy to explain, and it is a better calculation of most consistent achievement. At the same time, only informed professional judgement

can be used to weigh more recent achievement, and this application is much more difficult to explain.

Educators also need to be careful not to resort to normative referencing. Frary, Cross and Weber (1993) and McMillan (2001), found examples of normative referencing within criterion systems. This study showed that normative referencing helped participants establish a framework for achievement. However, teachers must be careful not to use it to determine overall grades as with all shortcuts presented by Heuristic Achievement.

5.3.7 Improving Teacher Moderation

Teacher moderation has great potential to increase construct validity and decrease Heuristic Assessment. There is wide support for this interlocution practice in the literature (Gipps, 1994; Crooks, Kane and Cohen, 1996; Moss, 2003; Cumming and Maxwell, 2004; Harlen, 2005b; 2006; 2012a; Western, 2006; Cooksey, Freebody and Wyatt-Smith, 2007; Black, et al., 2010; Gardner, et al., 2010; Wyatt-Smith, Klenowski and Gunn, 2010). Teachers should not evaluate in isolation (Cizek, Fitzgerald and Rachor, 1995). Educators bring experiences, values, as well as their reflections on student achievement. Researchers such as Ebel and Frisbie (1991), Speck (1998) and Bishop (1992) wrote about the challenges faced by classroom teachers when evaluating as they must act as both judge and advocate. Consequently, there are technical, educational, philosophical, and ethical conflicts, as educators are not only evaluating the student but themselves as instructors. Moderation allows teachers to come together in an ideal

symbiosis to discuss their evaluations and establish a salutary consensus. Stiggins (2001b) made a point that there is a responsibility to apply what we know about assessment. However, *Growing Success* only mentioned moderation in passing as part of its section on formative assessment (p.39). Fortunately, the practice has been encouraged in subsequent documents (Ontario, 2013b). However, the participants confirmed that the term is widely used in schools, but there is little follow through. Therefore, administrations must do more to insist on the practice.

Since the evidence suggests teacher moderation can improve assessment dependability, it is something to which Ontario schools need to pay more attention. In order to do so, barriers to moderation must be addressed. First, there is the matter of finding the time to moderate (Stiggins and Bridgeford, 1985). Second, moderation does not guarantee consensus (Malone, DeLucchi and Long, 2004; Hutchinson and Hayward, 2005). How teachers understand quality assessment varies (Luke, 2003). Harlen (2005a) noted that teachers must ensure they are using the same vocabulary in order to communicate properly. It is best that teachers cite examples to back up their opinions, especially when they are basing their judgement on something outside of the standards. Wyatt-Smith, Klenowski and Gunn (2010) applied a mixed methods approach to teacher moderation of Queensland, Australia. In short, they found moderation is effective but not easily achieved. Groups need to work together to reach a consensus, and not everything is agreed upon. There will be cases were teachers are clearly ruling against standards because of their personal beliefs and values, a quality that was noted in the present study. Third, Harlen (2005b) also pointed out that moderation can just

become an adjustment exercise if not carried out properly. Fourth, Pollitt (2012) asserted that while comparing student work helps with the reliability of grades, it is important to ensure that the work is ultimately judged against content standards and not normative referencing.

5.3.8 Improving the Quality of Informed Professional Judgement Summary

This section of the chapter addressed pragmatic ways that schools and teacher education programs have helped improve the development of informed professional judgement in regards to determining a report card grade. It begins with informing the educator of the important aspects of evaluation, such as focusing on actual achievement and looking at past behaviour. Also, legacy issues such as normative referencing and the use of average must be replaced with what is mandated in policy: criterion referencing and calculations that truly show a student's most consistent achievement with special consideration for more recent achievement. Furthermore, teachers need to work together in moderation exercises to help ensure consistency. Obviously, there are challenges to change. Educators are set in their ways, and there are always time constraints. Nevertheless, it is essential that administrators work with educators to help promote positive assessment reform to increase the authenticity of report card grades.

5.4 Proposed Amendments to *Growing Success*

Ontario has undergone a generation of assessment reform. There have been many accomplishments, but some issues still need to be addressed. For

instance, *Growing Success* accentuated the need for professional judgement, but did not properly explain what this term meant. Fullan (2001, p.37) observed, "Painful unclarity is experienced when unclear innovations are attempted under conditions that do not support the development of the subjective meaning of the change." The data showed that policy could be interpreted numerous ways – including the continuation of outdated practices. Such actions could be explained in such a way as to be circumstantially aligned to policy, but shadowed practices must be stopped. Some clarifications to policy could help smooth over its rougher points, as well as improving organization. Moreover, these suggestions take school independence into consideration to allow educators to review the local situation for better assessment practices.

5.4.1 Comparison to Other Canadian Educational Assessment Policies

It should be noted that, although this study points out issues with *Growing Success* that require clarification and/or correction, Ontario is well ahead of its fellow provinces and territories in regards to a detailed assessment and evaluation policy. Similar to the pre-*Growing Success* situation in Ontario, other provinces spread out governing rules for classroom assessment over several documents. For example, Joint (1993) and Western (2006) are key assessment documents in western and northern Canada. Provinces such as Alberta and British Columbia have experimented with progressive evaluation reform, such as limiting the use of letter grades and levels to focus on formative feedback (Millar, 2014). Furthermore, both provinces have identified the need for an updated and unified assessment policy. Alberta reached this conclusion in 2009, but a revised policy

has yet to be released (Weber, et al., 2009). Likewise, British Columbia does not know when its new document will be available. Both New Brunswick and Saskatchewan, also without unified policies, have announced intentions to address this issue. These details are important to note as the delays demonstrate the difficulty of establishing a holistic assessment policy.

Prince Edward Island recently released a revised policy. However, *Prince Edward Island Department of Education and Early Childhood Development Assessment Policy* (2014) is a mere seven pages, only expanding on recommendations from 2005. The island province admits a more comprehensive policy is needed. In the meantime, there is much less explanation than provided in *Growing Success*. The document only lists off points of assessment practices teachers are to follow, with no mention of issues such as professional judgement. In 2010, Manitoba announced it would be revising its assessment policy effective mid-2012. However, it was delayed two years. The *Provincial Report Card Policy and Guidelines: Partners for Learning Grades 1 to 12* (Manitoba, 2014), reads much like *Growing Success*. For example, there are several references to the importance of professional judgement. Furthermore, the document made more definitive statements, such as "Grading is a complex process that requires a teacher's professional judgement. There is no single, prescriptive way to determine final grades… [grades are to be] accurate, meaningful, and consistent" (Manitoba, 2014, p.5). However, this document encounters the same problem as *Growing Success* of translating levels into percentages. Individual levels are still pooled together to form a percentage grade on the final report card. Moreover, the

percentage grades on the report cards are translated into letter grades on the

student's official transcript, further hindering transparency and meaning. A

comparison of Manitoba and Ontario policies shows that an assessment policy is

never complete and can always be further refined.

Note that the timeframe in these other provinces mirrors the development

of *Growing Success*, indicating that assessment reform was not an Ontario-only

issue but concerned educational systems across the country. Furthermore, the fact

that many years have surpassed without the establishment of a central assessment

policy is evidence of the challenges of creating and enforcing such a document.

Therefore, Ontario is fortunate to have launched *Growing Success*. The task now

is to make decisions regarding the next generation of assessment reform to make

further improvements to the education system.

5.4.2 Explaining Informed Professional Judgement

Since informed professional judgement is vital to an effective assessment

policy, instead of burying a definition in the Glossary and making detached

references, a discussion of the concept should be part of the Fundamental

Principles. It is not necessary to change the current definition, but simply expand

on its importance. The what, how, and why of professional judgement, and what

makes it informed, should be made clear. For instance, the policy could

acknowledge that it is not possible to outline every possible assessment scenario.

As a result, educators will need to rely on their training and experience, with the

guidance of administration, to reach certain decisions. Expanding the definition

could fit in well with the other Fundamental Principles of being fair, transparent, and equitable. Furthermore, a more clearly defined informed professional judgement could serve as a maxim to bring the policy together as a coherent document. When educators are asked to reflect on their decisions and how they reached a conclusion, they can strengthen their discretion and daily practice. This line of thinking would be superior to the current common occurrence of Heuristic Assessment.

5.4.3 Analogies and Fictional Case Studies

The current era of education reform in Ontario was established in the Royal Commission on Learning, entitled *For the Love of Learning* (Ontario, 1995; Anderson and Jaafar, 2003). Throughout this document, short fictional case studies were used to demonstrate ideas. In a draft version of *Growing Success*, a similar approach was used (Ontario, 2008). Likewise, Manitoba (2014) uses vignettes to illustrate good assessment practice. Including such anecdotes throughout the policy, as opposed to disembodied excerpts from assessment literature, could help identify axioms. The following are suggestions for what analogies and fictional case studies, included in the individual sections of the policy, could look like to assist administrators and teachers reflecting on their procedures and to develop best practices.

5.4.3.1 Sample Analogy: The Unreliable Scale

Growing Success provided details about how teachers should assess, but the message can get lost in the details. Therefore, including analogies within the chapters can established greater clarity. For example:

> *Imagine a weight scale that always weighs approximately five pounds light. One can stand on it multiple times in a matter of minutes, and the same weight will appear. However, based upon experiment with other scales, five pounds needs to be added to this scale in order to be accurate. Despite this known flaw, the scale can still be said to be dependable because of its consistency. Furthermore, when it appears the flaw can be corrected by adding five pounds, the weight can be referred to as reliable. However, there is still a validity issue: it is assumed that the weight is light, and five pounds needs to be added in order to be correct – but this it is an educated guess based on experience and experimentation. Nevertheless, the measurement error is recognized as being unavoidable, and the adjusted weight can still be considered trustworthy.*

This analogy helps to demonstrate the conflict between reliability and validity in student assessment. Assessment is ultimately a subjective exercise, and there is always some type of measurement error present; there is something slightly off about the scale because we are attempting to measure learning – which resists a scientific approach to measurement. Classroom teachers need to be more like the qualitative researcher who looks to validate findings as opposed to using reliability and validity in the traditional psychometric sense. By recording student results from products, conversations, and observations, educators can look consistency in achievement. Just as one's weight can fluctuate day-to-day, so can evidence of

learning. Moreover, just as one's recent weight should be given more attention than what was recorded several month's before, a student's most recent achievement deserves special consideration. In other words, a student's summative scores can be high and low, but it takes the informed professional judgement of an educator to weigh the quality of evidence from different points in time, especially towards what a student has achieved recently, in order to determine an overall grade. Calculations can help provide a transparent foundation for decisions, but in the end it is a teacher's evaluation expertise that confronts measurement error, triangulates evidence, and decides upon a grade that is fair and equitable.

5.4.3.2 Sample Fictional Case Study: Failure to Complete Homework

Fictional Case Studies, based on scenarios familiar to Ontario teachers, could help teachers reflect on policy and its intent. For instance, in order to help show how student behaviours can influence evaluations, there could be a scenario about homework:

> *Every weeknight, Chris, the classroom teacher, assigns a short homework assignment covering the day's material and/or setting up the next lesson. Chris is frustrated that most students do not attempt the homework. There is a weekly quiz, lightly weighted for assessment of learning, also covering this material. Although most students in the class pass this quiz, Chris has noted, and told the students numerous times, that those who do the homework regularly tend to do better on the quiz. Still, completion of homework remains a problem. When Chris is reviewing this situation as part of determining overall grades, what advice would you give the teacher?*

Since policy is not a textbook, the fictional case study does not necessarily have to be followed with precise answers. It is up to the reader to reflect on the policy, and decide what would be the best course of action based upon what is stated. In the case above, it would be important that Chris was distinguishing between homework and quizzes when determining the learning skills and the overall grade. The teacher would need to be careful if formative and summative evidence are being used to compliment one another and assessment actions are clearly communicated to the students. How to go about this process can be the subject of staff meetings, teacher moderation, and/or professional development days to reach a consensus on best practices for the school.

5.4.4. Revised Statement on Punitive Measures

Punitive measures, arguably the most polarizing issue in Ontario education, will continue to create debates between educators. This issue is a delicate matter. The current policy states that teachers may use punitive measures, but implicitly suggest they should not be a first resort. Participants in this study suggested sometimes it is necessary to apply the measures to get the attention of certain students. However, if a school is also using average, one zero could distort a student's true level of achievement, which is a violation of policy. Therefore, policy should continue to allow teachers to use their professional judgement on punitive measures, including stressing the other steps be used before such procedures are applied. Furthermore, educators would need to demonstrate what actions are also being taken in order to report accurately. It could be as easy as the erasure of a zero or late marks because better evidence of learning is obtained

through other means. For example, reflecting on evidence collected during formative circumstances could be more appropriate than a zero on a summative assignment. The key is to ensure that grades are always based on construct validity, and not an emotional reaction to student behaviour.

Rules governing punitive measures illustrate the dangers of a fixed approach to assessment. Consistency is important, but strict conformity undermines an evaluation system of student learning. Educators, when confronted with disagreeable guidelines may apply Heuristic Assessment. Consequently, consistency is undermined. Punitive measures should only be considered on a case-by-case basis guided by the informed professional judgement of the instructor. When punitive measures are applied, the rationale should be clearly explained to all parties involved as well as the desired outcome, i.e., a zero will remain until the student makes up for the work to demonstrate achievement of the curriculum expectations. Whether or not the replacement work warrants full marks is another example of where teacher moderation can add consistency to school evaluation practices.

Again, based upon the evidence collected in this study, it can be said that most students who do not hand in assignments, or are consistently late with work, are likely to be weak academically and/or to be experiencing external issues threatening achievement. Therefore, they will benefit from counseling and some leniency. An earned low grade is superior to a gift pass. Stronger students can also benefit from the situation if they are also permitted to hand in revised assignments to demonstrate greater skill on the expectations. However, all

students will need to negotiate these steps with the teacher to ensure the process is fair, transparent, and equitable to all.

5.4.5 Statement on Calculating Versus Determining Grades

Although *Growing Success* explained evaluation in terms of containing both calculations and professional judgement (Section 1.6.5.4.1), the analysis showed the continued dominance of a mathematical approach to evaluation (e.g., Section 3.2.2). As a result, the participants used Heuristic Assessment to adjust grades – as opposed to determining them – when they did not feel the math added up. This study makes the argument that Heuristic Assessment is not what informed professional judgement is meant to be; it should not be an after-the-fact issue that is done in the shadows. Instead, professional judgement should be as transparent as possible and allow students and other stakeholders to understand how a grade was determined in a manner that is consistent with the Fundamental Principles.

5.4.5.1 A Warning About the Use of Average

Growing Success did not advocate any particular calculation method for grades. Instead, it stated that teachers should use their professional judgement to determine a student's achievement. Unfortunately, because of the legacy of the use of average, a more definitive statement is needed. As previously stated, the use of average is not a proper approach to determining a student's most consistent achievement, especially if punitive measures are part of the equation. However, since the policy needs to be flexible, a ban on the use of average could be too

disruptive to assessment practices in Ontario. Therefore, policy should advise

teachers that caution must be used when applying average, or mean, to student

achievement. Many researchers have agreed that the use of average distorts

achievement as including all scores is not the same as determining consistency

(Wormeli, 2006a; Scriffiny, 2008; O'Connor 2009, 2010; Reeves, 2010; O'Connor

and Wormeli, 2011). Using a method that more accurately reports most

consistent, such as median or mode, in conjunction with triangulation of student

products, conversations, and observations, should be encouraged within policy

(Marzano, 2000; Guskey and Bailey, 2001; Reeves, 2010; Herbst and Davies,

2014).

A second edition of *Growing Success* should clarify the difference between

process, progress, and product should be noted. It is also an opportunity to note

the true difference between formative and summative assessment, i.e., how

evidence of learning is used. Multiple sample calculations and examples of the

triangulation of assessment data, perhaps in fictional case studies, could show

what is meant mathematically by most consistent and more recent achievement,

and how different approaches yield different results. This scenario would help

illustrate professional judgement to when determining a student's report card

grade, and invite further in-school discourse on what approaches would work best

in the local situation.

5.4.6 Procedures for Borderline Students

Growing Success touched upon borderline students, but it deferred the issue to individual interpretation. If the province is going to have acceptable consistency in assessment practices, then the issue of passing borderline students must be addressed more directly. Furthermore, the shadowed practice in Ontario of adjusting grades to award students a pass is so common, based on the participants in this study and the literature (Earl, et al., 2010; Simon, et al., 2010), that it is practically a shared practice. Although boards need to be able to apply the policy for their local circumstances, some basic guidelines for what constitutes a passing grade in a Secondary School course in the province need to be established.

The participants in this study demonstrated that observations and conversations took on a greater significance when it came to borderline students. Chances are such students have missing assignments, so there is less tangible evidence of achievement of the curriculum expectations. Still, using minimum grades for minimal evidence of learning can help justify a passing grade with construct validity (Carifio and Carey, 2009). Furthermore, the participants noted that punitive measures alone should not keep students from passing if pupils have provided evidence of learning in other ways. This is another opportunity for a fictional case study to show how evidence gained from formative assessment could be put to a summative use. It is not about showing favouritism to a particular student, but applying good assessment practices by recognizing that evidence of learning can be used for multiple purposes.

However, *Growing Success* should continue its emphasis on the learning process and the importance of assessment as and for learning. When students are engaged in daily lessons, which build skills for summative assessment, students will have a better understanding of what is expected of them and complete assignments on time. Therefore, the evidence gained from formative or summative circumstances should be consistent with one another, while also encouraging the fundamental principles of assessment and evaluation in Ontario. It is the responsibility of the teacher, administration, and appropriate support staff, to assist students who are struggling in the learning process to limit the number of students becoming "borderline" in the first place.

5.4.7 Final Report Card Procedures

Although there are many issues that administrators and teachers can negotiate to best serve students, there are other procedures that must be uniform in the interest of province-wide consistency. The second edition of *Growing Success* could highlight areas where the administration needs to uphold standards and ensure teacher compliance. Teachers should still have autonomy, but part of professional judgement is to recognize and follow guidelines; one cannot have a blithe disregard for proper practice. Referring to leadership for procedures is an essential part of sound assessment and evaluation. Assessment policy cannot be mercurial if a consistent approach is to take hold in the province. At the same time, assessment practices need to be malleable to suit the individual situation. As long as the local administration makes these processes clear, practices are likely to

be more efficient and teachers will not have to rely on Heuristic Assessment to solve dilemmas.

During the school year, department heads should talk to teachers about individual assignments. Educators should be able to demonstrate how the four learning categories are built into instruction. By building assessments with a balance of the learning categories in mind, evaluation can focus on achievement of the overall expectations. As previously discussed, when Ontario teachers need to include the learning categories and expectations on assignments, it can become confusing what the evidence of learning reflects. Consequently, participants engaged in Heuristic Assessment to determine a symbolic grade that they felt reflected what the student had accomplished – but could have a weak connection to the actual overall expectations. If teachers are encouraged to obtain multiple pieces of evidence for the overall expectations, the reliability and validity of student evaluations will be improved.

In regards to the final report card, administrators must discuss final report card procedures with teachers and check for a common understanding. The teacher should be able to explain the process in a manner similar to how they would explain it to a student, parent/guardian, or other stakeholder. Providing an explanation should not be regarded as policing professional judgement, but demonstrating transparency in evaluation. At the same time, administrators and teachers should be sensitive to steps in the assessment process, and recognize when certain actions are not consistent with the values stated in policy. The goal is not for all teachers in the province to evaluate in the exact same way; such an

325

objective is not practical. Since all classroom assessment inevitably contains measurement error, the aim is to increase assessment literacy and acceptable uniformity for all concerned parties, while improving reliability, validity, and consistency in grading.

5.4.8 Responsibilities of Educators

Individual educators have a responsibility to understand assessment policy, as proper implementation is ultimately up to them. They must also feel confident that they have the trust of, and empowerment from, the administration (Ravich, 2010). Current standards-based education in Ontario places a huge demand on the professionalism of the teacher. Assessment, instruction, and one's own philosophy of education must act as one (Shepard, 1989; 1995; 2000b; Airasian, 1991; Schafer, 1991; Stiggins, 1991; Cizek, 1993; Whittington, 1999; Daugherty, et al., 2012; Bye, 2015; Muslin, 2015). As previously mentioned, a second edition of *Growing Success* does not require a massive rewrite; it is more of a matter of clarifying certain issues. The following points concern teacher responsibilities that could benefit from additional information.

5.4.8.1 Improved Record Keeping

The participants in this study demonstrated that, be in on a computer or in manual notes, it is expected that records were kept on summative student products. Some respondents kept informal records, such as homework checks. However, even though both sets of records are divided into examples of summative and formative, they both relate to student products. Participants admitted that they

were most likely to use memory when it came to conversations and observations. Furthermore, the perception of conversations and observations, most likely in the form of consistent behaviours, could also form the reasoning for adjustments. As discussed in the literature review, if a teacher relies on memory, and/or reflects on student behaviours for evaluation, the likely result is construct irrelevance, i.e., hodgepodge grading. *Growing Success* did state that triangulation was a vital part of ensuring reliability and validity to a grade, but it did not discuss how to do so practically.

Since policy cannot be a textbook, a second edition of *Growing Success* should not bog down the reader with a precise discussion of psychometrics. At the same time, it should elaborate on how a teacher's professional judgement triangulates achievement. Triangulating achievement is similar to how a qualitative researcher looks at multiple forms of evidence, and decides how it all fits into a valid explanation. Therefore, it is important to collect records for conversations and observations, as well as student products. However, this requirement could sound intimidating; the learning process is incessant. Must teachers keep detailed records for both formative and summative assessment of student products, conversations, and observations for each student? This approach would not be feasible. Fortunately, brief notes, collected over the course of a semester, could be beneficial. As stated previously, a student's everyday performance is likely to be consistent, with the hope that it will improve over the course of the learning process. The key is to have regular tangible records that will serve a teacher better than memory.

Growing Success could make some simple suggestions on how to keep regular records. For example, notes could be made as part of a lesson plan debrief. Granted, this approach is also relying on short-term memory. However, it is the consistency of student achievement that is important. If a teacher misinterpreted one piece of evidence, a more reliable interpretation should occur over time. It is not the intention of this study to address everyday classroom assessment. That said, how an educator collects student products, observations, and conversations, on a daily basis is another important issue to address by administrators. This day-to-day data is essential to informed professional judgement when it comes to the final report card grade.

5.4.8.2 Instructions for Teacher Moderation

As previously mentioned, educators should make a sincere effort towards teacher moderation, and administrators should provide necessary support to allow educators to fulfill this vital task. As previously discussed, assessment is ultimately subjective and disposed to measurement error. However, teacher bias is an element that can be controlled in the interest of students and construct validity (Hoge and Butcher, 1984; Bennett, et al., 1993; Brookhart, 1994; Guskey and Bailey, 2001; Harlen, 2005a; Volante, 2006; Musial, et al., 2009). Again, policy would need to be carefully worded so as not to set unrealistic expectations for individual schools. Teacher moderation was mentioned in *Growing Success*, but the conversation needs to be expanded. The power of teacher moderation as an exercise in consistency and good assessment practice should be emphasized.

Granted, teacher moderation was encouraged in subsequent OME documents (Ontario, 2011; Ontario, 2013b). However, follow-up questions with the participants suggest that there is not always time in practice. A second edition of *Growing Success* could provide suggestions on how to incorporate moderation into regular practice. It is not that all educators must evaluate in the same way, as doing so would interfere with professional judgement and create other issues of conflict. Instead, teachers should share evaluation rationale. The feedback can be non-binding, but to hear the perspectives of others can influence assessment practices and increase the dependability of scores. After all, most educators learn how to evaluate on the job, so such instruction could be better engrained in a school's culture.

5.4.9 Proposed Amendments to *Growing Success* Summary

Ontario has made a clear commitment to using assessment and evaluation to drive the learning process (Ontario, 2010a; 2011; 2013). However, instructions on how actually to determine grades are left to the interpretation of individual schools. The analysis demonstrates a disconnect in how to take evidence of learning and use it to inform professional judgement. Until this matter is corrected, the authenticity of Ontario final grades will be in question as educators engage in Heuristic Assessment. A revised edition of *Growing Success* can make instructions clearer, while still allowing schools to address their individual needs. By infusing informed professional judgement with construct validity, assessment reform in Ontario will continue to move in a positive direction.

5.5. Proposed Amendments to Reporting Student Learning

It is one matter to look through the existing policy and point out how it could be made clearer, or to expand on existing ideas. However, this study also encourages the OME to espouse further assessment reform. It is believed that the following corollaries are in line with the direction the province has been heading. Nevertheless, these steps would require modifications of policy. Furthermore, they would likely generate resistance. A revised edition of *Growing Success*, in addition to the previously discussed clarifications, may present the opportunity to make further changes that would benefit student learning.

5.5.1 Course Work and the Final Evaluation

A peculiar item that came up during the research was that informed professional judgement did not apply to the 30% final evaluation. Again, an Ontario Secondary Student's course grade is a dyad of course work and a final evaluation. A teacher must take whatever a student scored on the final evaluation, which may be divided into multiple pieces but is more likely to be one piece such as a final exam, and add it to the 70% of the course work that is comprised of a student's most consistent, with special consideration for more recent, achievement. Frankly, this is a counter-intuitive and convoluted approach. Since a reliable and valid evaluation should be based on multiple pieces of evidence, to lump 30% of evidence together and add it on without considering the context in which this evidence was obtained, does not make sense. Not surprisingly, this study found that participants, apparently with the blessing of their administrations, universally

ignored the directive. Airasian, Engemann and Gallagher (2007) would probably

agree that separating the 30% final evaluation from the rest of achievement would

not be appropriate as it breaks from the reality of the classroom. In a personal

correspondence, O'Connor (2014) expressed his demur over *Growing Success*,

specifically identifying the separation of the final evaluation from the course work

as a significant flaw. If a teacher's professional judgement is to be trusted, then it

needs to apply to the final evaluation as well. Furthermore, it should be strongly

suggested that this final evaluation to broken down into multiple pieces covering

all of the overall expectations, such as in an independent study project, to ensure

reliability and validity.

5.5.2 Improving the Consistency of Percentage Grades

Another oddity in Ontario Secondary Schools is the use of levels to score

individual student work, but the need to translate these levels into a percentage

grade on the report card. We have seen, as a result, the participants used Heuristic

Assessment to adjust grades, as opposed to using informed professional judgement

to triangulate. If the province is to continue using percentages along with levels, it

would be helpful to have a more precise conversion guide. Assuming that

administrations can make it clear what it means to establish a student's most

consistent level of achievement with special consideration for more recent

achievement, an expansion can be made to the chart that is currently found in

Growing Success (see Section 1.6.5.4.3.). Table 5.2: Revised Achievement Levels

and Corresponding Percentage Grades pegs a level to percentage grade. At the

same time, a range of variance is provided to allow for informed professional

judgement. An educator can further reflect on student products, conversations,

and observations, and can triangulate the percentage grade accordingly in a

transparent manner. For example, if a student's most consistent achievement

throughout the course was 4-, then that student has earned a percentage grade of

83%. The teacher can also reflect on more recent achievement and determine a

final grade of between 80%-86%. Of course, more recent achievement on heavily

weighted assignments could warrant an entirely different level. Whatever the

decision, the report card comments should specify the judgement was made.

Table 5.1: Revised Achievement Levels and Corresponding Percentage Grades

Level	-1	1	1+	2-	2	2+	3-	3	3+	4-	4	4+
Percentage	51	55	58	61	65	68	71	75	78	83	91	98
Triangulation Range	50-52	53-56	57-59	60-62	63-66	67-69	70-72	73-76	77-79	80-86	87-89	95-100

Note that Table 5.2, in order to remain consistent with what is currently stated in

Growing Success, only addresses percentages that correspond to passing grades.

5.5.3 Replacing Percentage Grades with Levels

However, a better approach to the Ontario report card would be to do away

with the percentage grade and only report the achievement level. Cooper (2007;

2011) has argued that percentage grades have more to do with politics than

pedagogy. It relates back to public misunderstanding of assessment, with the

possibility of credulity turning into backlash against the education system.

According to Cooper, the assumption is parents and other stakeholders understand

percentages and do not understand levels; therefore, the report card remains as a

percentage – even though levels are found in everyday classrooms. It is also a

legacy of the normative-referenced system. If Ontario is going to make a true

transition to a criterion-referenced system, then levels must replace percentages on

report cards (Cooper, O'Connor and Wakeman, 2009). O'Connor and Wormeli

(2011) pointed out that the popular International Baccalaureate program uses

seven levels of achievement on report cards while the American Advanced

Placement classes uses five. Therefore, the use of levels on report cards is not

without precedence. Although these other programs are more associated with

university-bound students, it does not mean that a levels-based evaluation cannot

communicate achievement for college- and workplace-bound students as well.

5.5.4 Evaluation of Individual Overall Expectations and the Report Card

Since a percentage grades combines all the curriculum expectations into

one score, it interferes with reliability (Frary, Cross and Weber, 1993). It is

difficult to ascertain the strengths and weaknesses of a student. Consequently, the

percentage grade does not communicate learning as an evaluation should.

Furthermore, the inability to interpret the message can result in the inability to

identify where a student can improve. If the purpose of assessment is to improve

student learning, as *Growing Success* professed, then percentage grades fail in this

endeavor.

However, what is an adequate solution? It would become far too

complicated to enumerate individual overall curriculum expectations; a typical

course has over a dozen such expectations. If the meaning of evaluation is lost in

percentages grades, creating an overly complicated system that provides too much information would communicate learning – but the message would be lost in numerical chaos. If a unified grade provides too little information, and evaluating many different aspects of learning is too much detail, is there a sufficient middle ground?

Since the Ontario curriculum already divides the overall expectations into multiple strands, each one of these groupings could be assigned a level. The report could state a cumulative level for the course overall which would have a mathematical basis. For example, if a student has achieved a range of 3- to 3+ on individual overall expectations, it could make sense to assign a cumulative Level 3. The report card comments could include additional information on how this cumulative level was determined. Thus the report card would communicate achievement with greater transparency and improve construct validity.

5.5.5 Modification of Levels

In regards to the levels used for summative assessment, further reform is required. The study's participants expressed confusion over what the levels meant, particularly the 20% range for Level 4, which represents exceeding the provincial standards. Explanations of why this range is twice as large as the 10% incremental Levels of 1-3 tend to be glib. This confusion was one of the major motivators for Heuristic Assessment.   A grade should clearly communicate student achievement; there should be no misperception. Therefore, a correction is needed for the levels. This study proposed a revised seven-level system. Levels

1-5 would include minuses and pluses to provide a range of variance for better accuracy and the ability to communicate student achievement, leaving the chart with eighteen possible levels and codes. Table 5.3: Proposed Level Grade Definitions demonstrates these proposed levels and their meaning, based on a table that is currently found on Ontario report cards (Ontario, 2010a, p.129). Granted, there is still a degree of subjectivity between the levels, hence the need for informed professional judgement to identify matters such as exceeding the provincial expectations and demonstrating mastery. It is hoped that when Ontario teachers are working exclusively in levels, and receive more professional development on the matter, they will better hone their discretion and recognize the consistent level of achievement.

Table 5.2: Proposed Level Grade Definitions

Level Grade	Achievement of the Provincial Curriculum Expectation
5	The student has demonstrated mastery of the required knowledge and skills with a high degree of effectiveness. Achievement has well surpassed the provincial standards.
4	The student has demonstrated the required knowledge and skills with a high degree of effectiveness. Achievement has surpassed the provincial standards.
3	The student has demonstrated the required knowledge and skills with a considerable degree of effectiveness. Achievement has met the provincial standards.
2	The student has demonstrated the required knowledge and skills with some effectiveness. Achievement is approaching the provincial standards.
1	The student has demonstrated the required knowledge and skills with limited effectiveness. Achievement falls well below the provincial standards.
R	The student has not demonstrated the required knowledge and skills for this course. Extensive remediation is required.
I	Insufficient evidence to assign a level grade.
W	The student has withdrawn from the course.

Also note that the option to assign the code of "I" for insufficient evidence has been extended to Grades 11 and 12. This change would require a policy modification as this code can only be used up to Grade 10. A percentage grade for Grade 11 and Grade 12 must be reported in the interest of full disclosure (Ontario,

2013c). In order to add greater consistency for all of Secondary School, the option should be available for senior students. Again, the code of "I" is not a passing grade, but indicates that there is not enough evidence to justify the failing code of "R." In regards to full disclosure, the fact that there was not enough evidence to even grant a code of "R" should be enough disclosure in itself.

In order to have a uniform report card, it would probably be too challenging to have a bespoke report of an individual student's classes with overall expectations printed in the matrix. Therefore, a means to communicate this information, such as a printout included with the report card would need to be provided for students, parents, and other stakeholders. In the case of CHY4U (a Grade 12 University-Stream History course entitled World History since the Fifteenth Century; see Ontario, 2015, pp.389-408) the overall strands and expectations are stated on Table 5.3a: Overall Expectations by Strands for CHY4U.

Table 5.4b: Sample Report Card Entry for CHY4U demonstrates how the student's achievement could appear on the final report card. Generally speaking, the overall expectations and learning categories should be treated with appropriate balance when the teacher is determining an overall grade. Table 5.4b attempts to show how the commutative level was based on a dynamic judgement, and not arbitrary standards (Broadfoot, 2002). It is hoped that the enlarged font used for the cumulative level allows the achievement to stand out, while the other details

Table 5.3a: Overall Expectation by Strands for CHY4U

Strands	Expectations
A: Methods of Historical Inquiry and Skill Development	A1: Historical Inquiry A2: Developing Transferable Skills
B: The World, 1450-1650	B1: Social, Economic, and Political Context B2: Communities, Conflict and Co-operation B3: Identity, Citizenship, and Heritage
C: The World, 1650-1789	C1: Social, Economic, and Political Context C2: Communities, Conflict and Co-operation C3: Identity, Citizenship, and Heritage
D: The World, 1789-1900	D1: Social, Economic, and Political Context D2: Communities, Conflict and Co-operation D3: Identity, Citizenship, and Heritage
E: The World Since 1900	E1: Social, Economic, and Political Context E2: Communities, Conflict and Co-operation E3: Identity, Citizenship, and Heritage

Table 5.3b: Sample Report Card Entry for CHY4U

Course	Commutative Level	Overall Expectation A		Overall Expectation B			Overall Expectation C			Overall Expectation D			Overall Expectation E			Final Evaluation Level
Course Title: World History since the Fifteenth Century **Course Code:** CHY4U **Teacher Name:** John Teacher	**4**	1	2	1	2	3	1	2	3	1	2	3	1	2	3	
		3	4-	4	5-	5	3	3	3	3+	4	4	4	4	4	
		3+		5			3			4			4			**4**

Learning Skills and Work Habits		Teacher Comments
Skills or Work Habit	**Evaluation**	
Responsibility	G	Teacher will insert a paragraph proving an overview of the student's learning process and explanation for how the various levels determined the Commutative Level of a Level 4 for most consistent achievement with special consideration for more recent, including the Final Evaluation.
Organization	G	
Independent Work	E	
Collaboration	G	
Initiative	E	
Self-Regulation	E	

are stated in the interest of transparency. The teacher comments should articulate

how the scores on the various strands contributed to the product. Again, detailed

records should be kept on various student products, conversations, and

observations covering the various strands and expectations to help make this

decision and justify it to third parties. An item that has been omitted from this

example report card is the class median. Comparing the student's product to

others overshadows the learning process and is another example of the normative legacy. Besides, the median would add more information to a report that aims to be transparent and avoids unnecessary clutter.

5.5.6 Proposed Amendments to Reporting Student Learning Summary

This study found that because of the confusing nature of assessment and evaluation, participants applied Heuristic Assessment to final report cards. As a result, it is difficult to make the claim the final grades in Ontario are truly consistent and can be considered as a reliable and valid communication of student achievement. The main culprit is the fact teachers are to assign levels, based on learning categories and overall expectations, to individual assignments – but then to convert all this information to a single grade based on a 100-point scale. Consequently, it is difficult to interpret a student's true overall achievement, and how s/he can improve; percentage grades do not improve student learning. Ontario could learn from other systems such as the International Baccalaureate to use levels for assignments on the report card. With some increased public awareness, it can be shown that breaking down the report card to illustrate the sources of how an overall level was determined, would not only make more sense, but also imbue transparency into evaluation. The next step would be to decide the best way to structure a revised report card, and whether or not the system is applicable to all academic subjects.

5.6 Weaknesses in the Study

The insight provided by this study is valuable; it is a look at significant concerns regarding assessment practices in Ontario, and contains thoughtful

recommendations on how to correct these issues on the basis of current research. However, no academic research can claim to be infallible. The Methodology and Methods chapters went to great lengths to show how steps were taken to establish a sound analysis. At the same time, a CGT opens itself to criticism as any report based on the voices of participants is but one possible view (see Section 2.3.8). Nevertheless, like measurement error in classroom assessment, there will always be unresolved issues that must be accepted in order to keep the conversation going.

There are also several issues that the study mentioned in passing but did not explore in detail. For example, if all assessment should improve student learning, how does a final report card grade fit into that concept? Granted, it could be linked to the next grade and life-long learning, but how can this success be connected to a final grade? It is assumed having more accurate reporting is a good start, but this issue should be explored in more detail. Also, it was decided not to pursue an analysis of the meaning of learning skills and work habits as the study focused on percentage grades, and perceptions of student behaviour were captured as part of Heuristic Assessment. However, what is the connection between the evaluation of the learning skills and work habits and the overall percentage grade? Could this information shed further light on Heuristic Assessment and improved professional judgement? On another note, in regards to how much leeway educators are given to apply informed professional judgement, is there a connection to the socio-economic situation of the school? Although this study provided insight into the application of professional judgement, the field of classroom evaluation is so vast that there will always be more issues to explore.

5.6.1 Possible Testing of Findings

CGT does not in itself prove findings via testing. Grounded theory is generally not concerned with such verification, as it is focused on generating hypotheses (Dey, 1999). Nevertheless, in order to solidify systemic claims about the nature of report card grading in Ontario, the present findings could be further investigated to determine the trustworthiness of the conclusions and proposals. For instance, the conversations with participants focused on their explanation of the report card process. As a result, they looked back on individual assignments, how they address common issues such as late and missing assignments (i.e., how punitive measures were applied), and how they view student achievement as a whole. Because Ontario's standards-based approach involves weighing overall curriculum expectations along with learning categories, a more comprehensive case study could have followed teachers through a course, examining the design and scoring of summative assessments, and discussions on how conversations and observations were included as part of informed professional judgement. Even though grounded theory could still be used with this kind of data gathering, it would have been much more demanding on the participants. They would have had to commit to frequent meetings and detailed discussions. Furthermore, there may have been site access issues as respondents would have had to share a lot of information on specific students, which would need approval from the local administrations to satisfy ethical concerns. In addition, observing the teachers in the classroom would have been needed to gain insight on how they interacted with

students. Still, conducting such a case study in the future would serve as an interesting comparison to the present findings.

5.6.2 Challenges to Implementing Recommendations

In regards to implementing recommendations made in this chapter, it may be said that they do not take into account objections from the various Teacher Unions, who may not greet further policy change with an effusive welcome. Additional record keeping would likely increase a teacher's workload and is bound to encounter opposition. Furthermore, there are both semester and non-semestered Secondary Schools in Ontario, not to mention four major boards, which present additional challenges to how the recommendations should be carried out. However, if the recommendations are designed to improve student learning, and to increase the efficiency of assessment and evaluation practices, it should be pursued. Discussing how these plans would benefit both teachers and students is a conversation worth having.

5.6.3 Possible Follow-Up Studies

Instead of looking at weaknesses in this study as flaws, it is better to look at them as opportunities for follow-up studies. The discussions started in this study have many possibilities. For example, it was already mentioned that teachers of math and science view the final report card process through a different lens than the participants in this study. How do they view the application of professional judgement on final report cards? Taking an approach to their experiences similar to the approach in this study could help to determine if Ontario would benefit from a multi-tiered system for different academic subjects.

Furthermore, because there is a significant divide in how elementary and secondary school students are evaluated, and the social consequences involved, it would make sense to investigate different editions of *Growing Success* for the two major levels of schooling. The two could provide enough overlap to emphasize the obvious connections, but with different instructions regarding the use of learning skills and the assigning of the final grade.

This study also focused on the views of teachers, including their perception of the administration. What about the perception of teachers by the administrators? Alternatively, what about the voices of Ministry supervisors? How have they explained final report card processes, and how would their explanations compare to the results in this study? A CGT project would also work for this type of research. Having a better sense of assessment and evaluation from an elevated point of view would have a valuable application in a revised edition of *Growing Success*.

5.7 Conclusion

Heuristic Assessment is a reaction to the realities of the classroom. It is a construct irrelevance created by certain assessment practices. An improved explanation of good assessment practices to educators, guided by a thoughtful administration, can reduce measurement error. More importantly, improving the transparency of assessment and evaluation in Ontario, such as confronting shadowed practices, could significantly improve the overall system. Nothing should be ineffable in regards to student assessment. Informed professional

judgement, in regards to a methodical assessment practice, needs to be explored

with teachers to help them understand its effectiveness, and how to explain it to

students and other stakeholders. When the assessment practice is more open, it

will be easier for all parties involved to recognize the steps involved and how

decisions are reached. Assessment will continue to be one of the major challenges

of education, but further exploring professional judgement will provide an

opportunity to make it more effective.

This study provides a comprehensive view of professional judgement and

reporting student learning in Ontario by providing an exegetical look at policy,

letting the voices from the participants narrate the various processes, and reviewed

the literature on the topic. A detailed account of the methodology and methods

used to bring together the data was also provided. The analysis was rigourously

checked, to aver the claims made. By providing insight into how the participants

apply professional judgement to final report cards in Ontario, this study has made

a contribution to the field of pedagogical knowledge.

REFERENCES

Aaronsohn, E., Holmes, J. E., Foley, T. and Wallowitz, J., 1994. *"Teacher-pleasing", traditional grading – and learning? A collaborative qualitative study.* New Britain, CT: Central Connecticut University.

Airasian, P.W., 1991. Perspectives on measurement instruction. *Educational Measurement: Issues and Practice*, 10(1), pp.13-16, 26.

Airasian, P.W., 1994. *Classroom assessment*. 2nd ed. New York: McGraw-Hill.

Airasian, P.W., 1996. *Assessment in the classroom*. New York: McGraw-Hill.

Airasian, P.W., Engemann, J.F. and Gallagher, T.L., 2007. *Classroom assessment: concepts and applications*. Singapore: McGraw Hill.

Airasian, P.W. and Jones, A.M., 1993. The teacher as applied measurer: realities of classroom measurement and assessment. *Applied Measurement in Education*, 6(3), pp.241-254.

Allal, L., 1988. Quantitative and qualitative components of teachers' evaluation strategies. *Teaching and Teacher Education,* 4(1), pp.41–51.

Allal, L., 2013. Teachers' professional judgement in assessment: a cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy and Practice*, 20(1), pp.20-34.

Allard, C.C., Goldblatt, P.F., Kemball, J.I., Kendrick, K., Millen, J. and Smith, D.M., 2007. Becoming a reflective community of practice. *Reflective Practice*, 8(3), pp.299-314.

Alvesson, M., 2011. *Interpreting Interviews*. Singapore: Sage.

Anders, P., and Richardson, V., 1992. Teacher as game-show host, bookkeeper, or judge? Challenges, contradictions, and consequences of accountability. *Teachers College Record*, 94(2), pp.382-397.

Anderson, S.E., and Jaafar, S.B., 2003. *Policy trends in Ontario education: 1990-2003*. Toronto: Ontario Institute for Studies in Education.

Angen, M.J., 2000. Evaluating interpretive inquiry: reviewing the validity debate and opening the dialogue. *Qualitative Health Research*, 10(3), pp.378-395.

Annells, M., 1996. Grounded theory method: philosophical perspectives, paradigm of inquiry, and postmodernism. *Qualitative Health Research*, 6(3), pp.379-393.

Annells, M., 1997a. Grounded theory method, part I: within the five moments of qualitative research. *Nursing Inquiry*, 4(2), pp.120-129.

Annells, M., 1997b. Grounded theory method, part II: options for users of the method. *Nursing Inquiry*, 4(3), pp.176-180.

Annells, M., 2006. Triangulation of qualitative approaches: hermeneutical, phenomenology and grounded theory. *Journal of Advanced Nursing*, 56(1), pp.55-61.

Artuso, A., 2014. EQAO scores have value: Education Minister Liz Sandals. *Toronto Sun*, [online] 2 March. Available at: <http://www.torontosun.com/2014/03/01/eqao-scores-have-value-education-minister-liz-sandals> [Accessed 24 September 2014].

Bailey, B., 2000. The impact of mandated change on teachers. In: N. Bascia and A. Hargreaves, eds. 2000. *The sharp edge of educational* change. London: Routledge Falmer.

Baker, E. and O'Neil, H., 1994. Performance assessment and equity: a view from the USA. *Assessment in Education: Principles, Policy and Practice*, 1, pp.11-26.

Banister, P., Burman, E., Parker, I., Taylor, M. and Tindall, C., 1994. *Qualitative methods in psychology: a research guide.* Buckingham: Open University Press.

Barnes, S., 1985. A study of classroom pupil evaluation: the missing link in teacher education. *Journal of Teacher Education,* 36(4), pp.46-49.

Beach, W., 1992. *Ethical education in American public schools.* West Haven, CT: National Education Association.

Bennett, R.E., 2011. Formative assessment: a critical review. *Assessment in Education: Principles, Policy and Practice*, 18(1), pp.5-25.

Bennett, R. E., Gottesman, R. L., Rock, D. A. and Cerullo, F., 1993. Influence of behaviour perceptions and gender on teachers' judgements of students' academic skill. *Journal of Educational Psychology,* 85, pp.347–356.

Bishop, J.H., 1992. Why U.S. students need incentives to learn. *Educational Leadership*, 49(6), pp.15-18.

Birks, M., Chapman, Y. and Francis, K., 2008. Memoing in qualitative research: probing data and processes. *Journal of Research in Nursing*, 13(1), pp.68-75.

Birks, M. and Mills, J., 2011. *Grounded theory: a practical guide*. Singapore: Sage.

Birks, M., Mills, J., Francis, K. and Chapman, Y., 2009. A thousand words paint a picture: the use of storyline in grounded theory research. *Journal of Research Nursing*, 14(5), pp.405-417.

Black, P., 1993. Formative and summative assessment by teachers. *Studies in Science Education,* 21, pp.49–97.

Black, P., 2003. Testing, testing: listening to the past and looking to the future. *Social Science Review*, 85(311), pp.69-77.

Black, P., 2015. Formative assessment – an optimistic but incomplete vision. *Assessment in Education: Principles, Policy, and Practice*, 22(1), pp.161-177.

Black, P., Harrison, C., Hodgen, J., Marshall, B. and Serret, N., 2010. Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy and Practice*, 17(2), pp.215-232.

Black, P. and Wiliam, D., 2003. In praise of educational research: formative assessment. *British Educational Research Journal*, 29(5), pp. 623-637.

Black, P. and Wiliam, D., 2012. The reliability of assessments. In: J. Gardner, ed. 2012. *Assessment and learning*. 2nd ed. Singapore: Sage, pp.243-263.

Black-Meddings, L., Grebenc, B., Houlahan, K., MacMaster, M., Newstead, S., Page, M., Traxler, D., Elliott, P., Bischof, H., and Robillard, M., 2010. *Growing success policy workgroup report*. Toronto: Ontario Secondary School Teachers Federation.

Blaxter, L., Hughes, C. and Tight, M., 2006. *How to research*. 3rd ed. Maidenhead, England: Open University Press.

Blumer, H., 1969. *Symbolic interactionism: perspective and method*. 1986. Berkley: University of California Press.

Bogdan, R., and Biklen, S., 2007. *Qualitative research for education: an introduction to theories and methods*. 5th ed. Singapore: Pearson International Education.

Bond, L. 1995. Unintended consequences of performance assessment: issues of bias and fairness. *Educational Measurement: Issues and Practice*, 14(4), pp.21-24.

Bowen, G.A., 2008. Naturalistic inquiry and the saturation concept: a research note. *Qualitative Research*, 8(1), pp.137-152.

Bowers, B. and Schatzman, L., 2009. Dimensional analysis. In: J.M. Morse, P. Noerager Stern, J. Corbin, B. Bowers, K. Charmaz and A.E. Clarke, eds. 2009. *Developing grounded theory: the second generation*. Walnut Creek, CA: Left Coast Press, pp.86-127.

Boychuk-Duchscher, J.E. and Morgan, D., 2004. Grounded theory: reflections on the emergence vs. forcing debate. *Journal of Advanced Nursing*, 48(6), pp.605-612.

Brantlinger, E.A., 1993. *The politics of social class in secondary school: views of affluent and impoverished youth.* New York: Teachers College Press.

Broadfoot, P., 2002. Dynamic versus arbitrary standards: recognizing the human factor in assessment. *Assessment in Education: Principles, Policy and Practice*, 9(2), pp.157-159.

Brookhart, S.M. 1991. Grading practices and validity. *Educational Measurement: Issues and Practice*, 10(1), pp.35-36.

Brookhart, S.M. 1993. Teachers' grading practices: meaning and values. *Journal of Educational Measurement*, 30, pp.123-142.

Brookhart, S.M. 1994. Teachers' grading: practice and theory. *Applied Measurement in Education*, 7(4), pp.279-301.

Brookhart, S.M. 1997. A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education*, 10(2), pp.161-180.

Brookhart, S.M. 1999. Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practice*, 18(1), pp.5-13.

Brookhart, S.M., 2001. Successful students' formative and summative assessment use of assessment information. *Assessment in Education*, 8(2), pp.153-169.

Brookhart, S.M., 2004. *Grading*. Upper Saddle River, NJ: Pearson Education.

Brookhart, S.M., 2011. Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, 30(1), pp.3-12.

Brookhart, S.M. and Freeman, D.J., 1992. Characteristics of entering teacher candidates. *Review of Educational Research*, 10(1), pp.37-60.

Brown, G.T.L., 2004. Teachers' conceptions of assessment: implications for policy and professional development. *Assessment in Education: Principles, Policy and Practice*, 11(3), pp.301-318.

Brown, G.T.L., Glasswell, K. and D. Harland, D., 2004. Accuracy in the scoring of writing: studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, 9(2), pp.105–121.

Bryant, A., 2002. Re-grounding Grounded Theory. *The Journal of Information Technology Theory and Application*, 4(1), pp.25-42.

Bryant, A., 2003. A constructive/its response to Glaser's: constructivist grounded theory? *Historical Social Research*, 19, pp.106-113.

Bryant, A., 2007. *Research training 2-day workshop – grounded theory*. [online] Leeds Metropolitan University. Available at: <http://www.youtube.com/watch?v=fmWKf5L0mfA&list=PL8CB91CC62C1C2C 7E > [Accessed 5 November 2013].

Bryant, A., 2009. Grounded theory and pragmatism: the curious case of Anselm Strauss. *Forum: Qualitative Social Research*, 10(3), Art. 2. [online] Available at: <http://www.qualitative-research.net/index.php/fqs/article/viewArticle/1358> [Accessed 19 October 2014].

Bryant, A. and Charmaz, K., 2007a. Grounded theory research: methods and practices. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.1-30.

Bryant, A. and Charmaz, K., 2007b. Grounded theory in historical perspective: an epistemological account. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.31-57.

Bryman, A., 2004. *Social Research Methods*. 2nd ed. Oxford: Oxford University Press.

Buckley, C.A. and Waring, M.J., 2013. Using diagrams to support the research process: examples from grounded theory. *Qualitative Research*, 13(2), pp.148-172.

Bye, J., 2015. Foucault and the use of critique: breaching the self-evident of educational practices. *International Journal of Qualitative Studies in Education*, 28(4), pp.394-414.

Caelli, K., Ray, L. and Mill, J., 2003. 'Clear as mud': toward greater clarity in generic qualitative research. *International Journal of Qualitative Methods*, 2(2), pp.1-13.

Calman, L., 2011. What is grounded theory? *YouTube*. [online] Available at: <http://www.youtube.com/watch?v=AgQltbyUYik&list=PL8CB91CC62C1C2C7 E> [Accessed 19 February 2013].

Canady, R.L. and P.R. Hotchkiss. 1989. It's a good score: just a bad grade. *Phi Delta Kappa*, September, pp.68-71.

Cannella, G.S. and Lincoln, Y.S., 2011. Ethics, research regulations, and critical social science. In: N.K. Denzin and Y.S. Lincolin, eds. 2011. *The Sage handbook of qualitative research*, 4th ed. Singapore: Sage, pp.81-90.

Carifio, J. and Carey, T., 2009. A critical examination of current minimum grading policy recommendations. *The High School Journal*, 93(1), pp.23-37.

Carpenter, J., 2008. Metaphors in qualitative research: shedding light or casting shadows? *Research in Nursing and Health*, 31(3), pp.274-282.

Castellani, B., 1999. Michel Foucault and Symbolic Interactionism. *Studies in Symbolic Interaction*, 22, pp.247-272.

Chamberlain-Salaun, J., Mills, J. and Usher, K., 2013. Linking symbolic interactionism and grounded theory methods in a research design: from Corbin and Strauss' assumptions to action. *Sage Open*, July-September, pp.1-10.

Charmaz, K., 1980. *The social reality of death: death in contemporary America*. Reading, MA: Addison-Wesley.

Charmaz, K., 1983. The grounded theory method: an explication and interpretation. In: R.M. Emerson, ed. 1983. *Contemporary field research*. Boston: Little and Brown, pp.1161-1172.

Charmaz, K., 1990. Discovering chronic illness: using grounded theory. *Social Science and Medicine,* 30(11), pp.1161-1172.

Charmaz, K., 1995a. Between positivism and postmodernism: implications for methods. *Studies in Symbolic Interaction*, 17, pp.43-72.

Charmaz, K., 1995b. Grounded theory. In: J.A. Smith, R. Harre and L. van Longenhove, eds. 1995. *Rethinking methods in psychology*. London: Sage, pp.27-29.

Charmaz, K., 2000. Grounded theory: objectivist and subjectivist methods. In: N.K. Denzin and Y.A. Lincoln, eds. 2000, *Handbook of qualitative research*. 2nd ed. London: Sage, pp.509-35.

Charmaz, K., 2001. Grounded theory in ethnography. In: P. Atkinson, A. Coffey, S. Delamont, J. Lofland and L. Lofland, eds. 2001. *Handbook of ethnography*, Thousand Oaks, CA: Sage, pp.160-174.

Charmaz, K., 2004. Grounded theory. In: S.N. Hesse-Biber and P. Leavy, eds. 2004. *Approaches to qualitative research: a reader on theory and practice*. Oxford: Oxford University Press, pp.496-521.

Charmaz, K., 2005. Grounded theory in the 21st century: applications for advancing social justice studies. In: N.K. Denzin and Y.S. Lincoln, eds. 2005. *The Sage handbook of qualitative research*. 3rd ed. New Delhi: Sage, pp.507-535.

Charmaz, K., 2006. *Constructing grounded theory: a practical guide through qualitative analysis*. Singapore: Sage.

Charmaz, K., 2008a. A future for symbolic interactionism. *Studies in Symbolic Interaction*, 32, pp.51-59.

Charmaz, K., 2008b. Constructionism and the grounded theory method. In: J.A. Holstein and J.F. Gubrium, eds. 2008. *Handbook of constructionist research*. New York: Gilford Press, pp.397-412.

Charmaz, K., 2008c. The legacy of Anselm Strauss for constructivist grounded theory. *Studies in Symbolic Interaction*, 32, pp.127-141.

Charmaz, K., 2009. Shifting the grounds: constructivist grounded theory methods. In: J.M. Morse, P. Noerager Stern, J. Corbin, B. Bowers, K. Charmaz and A.E. Clarke, eds. 2009. *Developing grounded theory: the second generation*. Walnut Creek, CA: Left Coast Press, pp.127-193.

Charmaz, K., 2010. Grounded theory as emergent method. In: S.N. Hesse-Biber and P. Leavy, eds. 2010. *Handbook of emergent method*. New York: Gilford, pp. 155-170.

Charmaz, K., 2011. Grounded theory methods in social justice research. In: N.K. Denzin and Y.S. Lincoln, eds. 2011. *The Sage handbook of qualitative research*. 4th ed. Singapore: Sage, pp.359-380.

Charmaz, K., 2012. The power and potential of grounded theory. [online] *2012 Medical sociology annual conference.* British Sociological Association. Available at: <http://www.youtube.com/watch?v=zY1h3387txo> [Accessed 19 February 2013].

Charmaz, K., 2014. *Constructing grounded theory*. 2nd ed. Thousand Oaks, CA: Sage.

Charmaz, K. and Belgrave, L.L., 2012. Qualitative interviewing and grounded theory analysis. In: J.F. Gubrium, J.A. Holstein, A.B. Marvasti and K.D. McKinney, eds. 2012. *The Sage Handbook of interview research*. Thousand Oaks, CA: Sage, pp.347-366.

Charmaz, K. and Bryant, A., 2010. Grounded theory and credibility. In: D. Silverman, ed. 2010. *Qualitative research*. 3rd ed. Thousand Oaks, CA: Sage, pp.291-310.

Charmaz, K. and Mitchell, R., 1996. The myth of silent authorship: self, substance, and style in ethnographic writing. *Symbolic Interaction*, 19(4), pp.285-302.

Charon, J.M., 2010. *Symbolic interactionism: an introduction, an interpretation, an integration*. 10th ed. Upper Saddle, NJ: Prentice Hall.

Chase, S.E., 2005. Narrative inquiry: multiple lenses, approaches, voices. In: N.K. Denzin and Y.S. Lincoln, eds. 2005. *The Sage handbook of qualitative research*. 3rd ed. New Delhi: Sage, pp.651-680.

Chellew, D., MkBk_Support@bell.net, 2013. *Markbook: bumping 9s*. [email] Message to G. Cavanaugh (gordon.cavan@taylors.edu.my). Sent Sunday October 13 2013: 14:56. [Accessed 16 October 2013].

Cherryholmes, C.H., 1988. Construct validity and discourses of research. *American Journal of Education*, 96(3), pp.421-457.

Cheung, D., 2002. Refining a stage model for studying teacher concerns about educational innovations. *Australian Journal of Education*, 46(3), pp.305-322.

Chiovitti, R.F. and Piran, N., 2003. Rigour and grounded theory research. *Journal of Advanced Nursing*, 44(4), pp.427-35.

Christians, C.G., 2011. Ethics and politics in qualitative research. In: N.K. Denzin and Y.S. Lincoln, eds. 2011. *The Sage handbook of qualitative research*. 4th ed. Singapore: Sage, pp.61-80.

Cizek, G.J., 1993. Some thoughts on educational testing: measurement policy issues into the next millennium. *Educational Measurement: Issues and Practice,* 12(3), pp.10-16.

Cizek, G.J., Fitzgerald, S.M. and Rachor, R.E., 1995. Teachers' assessment practices: preparation, isolation and the kitchen sink. *Educational Assessment*, 3(2), pp.159-179.

Clarke, A.E., 2003. Situational analysis: remapping grounded theory after the postmodern turn. *Symbolic Interaction*, 26(4), pp.553-576.

Clarke, A.E., 2005. *Situational analysis: grounded theory after the postmodern turn*. New Delhi: Sage.

Clarke, A.E., 2007. Grounded theory: conflicts, debates and situational analysis. In: W. Outhwaite and S. P. Turner, eds. 2007. *Handbook of social science methodology*. Thousand Oaks, CA: Sage, pp.423-442.

Clarke, A.E., 2009. From grounded theory to situational analysis: what's new? Why? How? In: J.M. Morse, P. Noerager Stern, J. Corbin, B. Bowers, K. Charmaz and A.E. Clarke, eds. 2009. *Developing grounded theory: the second generation*. Walnut Creek, CA: Left Coast Press, pp.194-235.

Clarke, A.E., adele.clarke@ucsf.edu, 2014. *A thank you and request for clarification*. [email] Message to G Cavanaugh (gordon.cavan@taylors.edu.my). Sent Thursday 6 March 2014: 23:44. [Accessed 7 March 2014].

Clarke, A.E. and Friese, C., 2007. Grounded theorizing using situational analysis. In: K. Charmaz and A. Bryant, eds. 2007. *The Sage handbook of grounded theory*. London: Sage, pp.363-397.

Clough, P. and Nutbrown, C., 2012. *A student's guide to methodology: justifying inquiry*. 3rd ed. Singapore: Sage.

Coffey, A. and Atkinson, P., 1996. *Making sense of qualitative data: contemporary research strategies*. London: Sage.

Cooksey, R.W., Freebody, P. and Wyatt-Smith, C.M., 2007. Assessment as judgement-in-context: analysing how teachers evaluate students' writing. *Educational Research and Evaluation*, 13(5), pp.401-34.

Cooper, D., 2007. *Talk about assessment: strategies and tools to improve learning.* Toronto, ON: Thomson Nelson.

Cooper, D., dcooper3@rogers.com, 2010. *Researching assessment and evaluation*. [email] Message to G Cavanaugh (kabx9gwc@nottingham.edu.my) Sent 5 March 2010: 12:00. [Accessed 5 March 2010].

Cooper, D., 2011. *Redefining fair: how to plan, assess, and grade for excellence in mixed-ability*. Bloomington, IN: Solution Tree Press.

Cooper, D., O'Connor, K., and Wakeman, N., 2009. Redefining fair: assessment and grading in the 21st century. *Changing Perspectives*, February, pp.27-31.

Corbin, J., 1991. Anselm Strauss: an intellectual biography. In: D. Maines, ed. 1991. *Social organization and social processes: essays in honor of Anselm Strauss*. New York: Aldine De Gruyter, pp.17-42.

Corbin, J., 1998. Alternative interpretation: valid or not? *Theory & Psychology*, 8(1), pp.121-128.

Corbin, J., 2009. Taking an analytic journey. In: J.M. Morse, P. Noerager Stern, J. Corbin, B. Bowers, K. Charmaz and A.E. Clarke, eds. 2009. *Developing grounded theory: the second generation*. Walnut Creek, CA: Left Coast Press, pp.35-53.

Corbin, J. and Strauss, A., 1990. Grounded theory research: procedures, canons and evaluative criteria. *Qualitative Sociology*, 13, pp.3-21.

Corbin, J. and Strauss, A., 2008. *Basics of qualitative research: techniques and procedures for developing grounded theory*. 3rd ed. Singapore: Sage.

Costa, A. and Kallick, B. 2001. *Describing 16 habits of mind*. Alexandria, VA: Association for Supervision and Curriculum Development.

Costello, C. and McKillop, B., 2000. Dealing with late and absences. *Orbit*, 30(4), pp.45-46.

Covan, E.K., 2007. The discovery of grounded theory in practice: the legacy of multiple mentors. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.58-74.

Creswell, J.W., 2008. *Educational research: planning, conducting, and evaluating quantitative and qualitative research*. 3rd ed. Singapore: Pearson Education.

Creswell, J.W., 2012. *Qualitative inquiry and research design: choosing among the five approaches*. 3rd ed. Los Angeles: Sage.

Creswell, J.W. and Miller, D., 2000. Determining validity in qualitative inquiry. *Theory Into Practice*, 39(3) pp.124-130.

Criminal Records Act 1985, s.6.3(Ch. 47), Ottawa: R.S.C.

Cronbach, L.J., 1989. Construct validation after thirty years. In: R.E. Linn, ed. 1989. *Intelligence: measurement, theory and public policy*. Urbana, IL: University of Illinois Press, pp.147-171.

Crooks, T.J., 1988. The impact of classroom evaluation practices on students. *Review of Educational Research,* 58(4), pp.438-481.

Crooks, T.J., Kane, M.T. and Cohen, A., 1996. Threats to the valid use of assessments. *Assessment in Education: Principles, Policy and Practice*, 3(3), pp.265-286.

Cross, L.H. and Frary, R.B., 1999. Hodgepodge grading: endorsed by students and teachers alike. *Applied Measurement in Education*, 12(1), pp.53-72.

Cumming, J.J. and Maxwell, G.S., 2004. Assessment in Australian schools: current practice and trends. *Assessment in Education: Principles, Policy and Practice*, 11(1), pp.89-108.

Cumming, J.J., Wyatt-Smith, C.M., Elkins, J. and Neville, M., 2006. *Teacher judgement: building an evidentiary base for quality literacy and numeracy education.* Brisbane: Queensland Studies Authority.

Cutcliffe, J.R., 2000. Methodological issues in grounded theory. *Journal of Advanced Nursing*, 31(6), pp. 1476-1484.

Cutcliffe, J.E., 2005. Adapt or adopt: developing and transgressing the methodological boundaries of grounded theory. *Journal of Advanced Nursing*, 51(4), pp.421-428.

Cutcliffe, J.R. and McKenna, H.P., 2004. Expert qualitative researchers and the use of audit trails. *Journal of Advanced Nursing*, 45(2), pp. 126-133.

Czarniawska, B., 2009. The uses of narrative in social science research. In: M. Hardy and A. Bryman, eds. 2010. *The Sage handbook of data analysis*. Singapore: Sage, pp.649-666.

Datnow, A. and Hubbard, L., 2015. Teachers' use of assessment data to inform instruction: lessons from the past and prospects for the future. *Teachers College Record*, 117, pp.1-26.

Daugherty, R., Black, P., Ecclestone, K., James, M. and Newton, P.E., 2012. Alternative perspectives on learning outcomes: challenges for assessment. In: J. Gardner, ed. 2012. *Assessment and learning*. 2nd ed. Singapore: Sage, pp.72-86.

Davies, A., 2000. *Making classroom assessment work*. Courtenay, B.C.: Connections Publishing.

Davies, A., 2007. *Making classroom assessment work.* 2nd ed. Courtenay, B.C.: Connections Publishing.

Davies, A., 2011. *Making classroom assessment work.* 3rd ed. Courtenay, B.C.: Connections Publishing.

Davison, K.G., 2006. Dialectical imagery and postmodern research. *International Journal of Qualitative Studies in Education*, 19(2), pp.133–146.

Delamont, S. and Atkinson, P., 2010. Qualitative research and the postmodern turn. In: M. Hardy and A. Bryman, eds. 2010. *The Sage handbook of data analysis*. Singapore: Sage, pp.667-682.

Denzin, N.K., 2007. Grounded theory of the politics of interpretation. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp. 454-471.

Denzin, N.K. and Lincoln, Y.S., 1994a. Entering the field of qualitative research. *Handbook of qualitative research*. In: N.K. Denzin and Y.S. Lincon, eds. 1994. Thousand Oaks, CA: Sage, pp.1-17.

Denzin, N.K. and Lincoln, Y.S., 1994b. Preface. *Handbook of qualitative research*. In: N.K. Denzin and Y.S. Lincoln, eds. 1994. Thousand Oaks, CA: Sage, pp.ix-xii.

Denzin, N.K. and Lincoln, Y.S., 2005. Introduction: the discipline and practice of qualitative research. In: N.K. Denzin and Y.S. Lincoln, eds. 2005. *The Sage handbook of qualitative research*. 3rd ed. New Delhi: Sage, pp.1-31.

Denzin, N.K. and Lincoln, Y.S., 2011. The discipline and practice of qualitative research. In: N.K. Denzin and Y.S. Lincolin, eds. 2011. *The Sage handbook of qualitative research*. 4th ed. Singapore: Sage, pp.1-25.

Dey, I., 1999. *Grounding grounded theory guidelines for qualitative inquiry*. San Diego: Academic Press.

Dey, I., 2007. Grounding categories. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.167-190.

Dilley, P., 2000. Conducting successful interviews: tips for intrepid research. *Theory into practice*, 39(3), pp.131-137.

Draucker, C.B., Martsolf, D.S., Ross, R. and Rush, T.B., 2007. Theoretical sampling and category development in grounded theory. *Qualitative Health Research*, 17(8), pp.1137-48.

Dressman, M., Journell, W. and Mann, J., 2012. Teacher education: qualitative research approaches. In: S. Delemont, ed. 2012. *Handbook of qualitative research in education*, pp.181-194.

Duncan, C.R. and Noonan, B., 2007. Factors affecting teachers' grading and assessment practices. *Alberta Journal of Educational Research*, 53(1), pp.1-21.

Earl, L., 1995. Assessment and accountability in education in Ontario. *Canadian Journal of Education*, 20(1) pp.45-55.

Earl, L.M., Katz, S., Herbert, M., Elgie, S., Jodouin, H. and Clark, R., 2010. *Classroom assessment in Ontario secondary schools: in teachers' hands*. Toronto: Ontario Teachers' Federation.

Ebel, R.L. and Frisbie, D.A., 1991. *Essentials of educational measurement*. 5th ed. Englewood Cliffs, NJ: Prentice Hall.

Elwood, J., and Klenowski, V., 2002. Creating communities of shared practice: assessment use in learning and teaching. *Assessment and Evaluation in Higher Education*, 27(3), pp.243-256.

Elwood, J. and Murphy, P., 2015. Assessment systems and cultural scripts: a sociocultural theoretical lens on assessment practice and products. *Assessment in Education: Principles, Policy, and Practice*, 22(2), pp.182-192.

Erickson, F., 2011. A history of qualitative inquiry in social and educational research. In: N.K. Denzin and Y.S. Lincolin, eds. 2011. *The Sage handbook of qualitative research*, 4th ed. Singapore: Sage, pp.43-59.

Fleer, M., 2015. Developing an assessment pedagogy: the tensions and struggles in re-theorizing assessment from a cultural-historical perspective. *Assessment in Education: Principles, Policy, and Practice*, 22(2), pp.224-246.

Flick, U., 2002. *An introduction to qualitative research*. 2nd ed. London: Sage.

Fontana, A. and Frey, J.H., 2005. The interview: from neutral stance to political involvement. In: N.K. Denzin and Y.S. Lincoln, eds. 2005. *The Sage handbook of qualitative research*. 3rd ed. New Delhi: Sage, pp.695-728.

Forsey, M., 2012. Interviewing individuals. In: S. Delemont, ed. 2012. *Handbook of qualitative research in education*, pp.364-376.

Frary, R.B., Cross, L.H. and Weber, L.J., 1993. Testing and grading practices and opinions of secondary teachers of academic subjects: implications for instruction in measurement. *Educational Measurement: Issues and Practice*, 12(3), pp.23-30.

Freebody, P. and Wyatt-Smith, C.M., 2004. The assessment of literacy: working the zone between 'system' and 'site' validity. *Journal of Educational Enquiry,* 5(2), pp.30-49.

Friedman, S. J. and Frisbie, D.A., 1995. The influence of report cards on the validity of grades reported to parents. *Educational and Psychological Measurement*, 55(1), pp.5-26.

Friedman, S. J. and Manley, M., 1992. Improving high school grading practices: "experts" vs. practitioners. *NASSP Bulletin*, 76(544), pp.100-104.

Friedman, S. J. and Troug, A. J., 1998. Evaluation of high school teachers' written grading policies. *ERS Spectrum*, 17(3), pp.34-42.

Frisbie, D.A. and Waltman, K.K. 1992. Developing a personal grading plan. *Educational Measurement: Issues and Practice*, 11, pp.35-42.

Fullan, M., 2001. *The new meaning of educational change*, 3rd ed. Teachers College Press Columbia, New York.

Fullan, M., 2012. What America can learn from Ontario's education success. *The Atlantic* [online] 4 May. Available at: <http://www.theatlantic.com/national/archive/2012/05/what-america-can-learn-from-ontarios-education-success/256654/> [Accessed May 5 2012].

Fullan, M., 2013. *Great to excellent: launching the next stage of Ontario's educational agenda*. Toronto: Ontario Ministry of Education.

Garcia, G.E., 1995. Equity challenges in authentically assessing students from diverse backgrounds. *Educational Forum*, 59(1), pp.64-73.

Gardner, J., 2012a. Quality assessment practice. In: J. Gardner, ed., 2012. *Assessment and learning*. 2nd ed. Singapore: Sage, pp.103-122.

Gardner, J., ed., 2012b. *Assessment and learning*. 2nd ed. Singapore: Sage.

Gardner, J., Harlen, W., Haywood, L., Stobart, G. and Montgomery, M., 2010. *Developing teacher assessment*. Maidenhead, England: Open University Press.

Geertz, C., 1973. Thick description: toward an interpretive theory of culture. In: C. Geertz, ed. 1973. *The interpretation of cultures*. New York: Basic Books, pp.2-30.

Gibson, B., 2007. Accommodating critical theory. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.436-453.

Gidney, R.D., 1999. *From Hope to Harris: the reshaping of Ontario's schools*. 2nd ed. Toronto: University of Toronto Press.

Gill, K., cap@ontario.ca, 2013a. *Assessment and evaluation practices*. [email] Message to G. Cavanaugh (gordon.cavan@taylors.edu.my) Sent Tuesday, September 24, 2013: 15:18. [Accessed September 24, 2013].

Gill, K., cap@ontario.ca, 2013b. *Secondary evaluation policy*. [email] Message to G. Cavanaugh (gordon.cavan@taylors.edu.my) Sent May 2, 2013: 14:14. [Accessed May 3, 2013].

Gipps, C., 1994. *Beyond testing: towards a theory of educational assessment*. London: Routledge-Falmer.

Gipps, C., 1995. What do we mean by equity in relation to assessment? *Assessment in Education: Principles, Policy and Practice*, 2(3), pp.271-281.

Glaser, B.G., 1978. *Theoretical sensitivity: advances in the methodology of grounded theory*. San Francisco: University of California Press.

Glaser, B.G., 1992. *Emergence vs. forcing: basics of grounded theory analysis*. Mill Valley, CA: Sociology Press.

Glaser, B.G., 1998. *Doing grounded theory: issues and discussion*. Mill Valley, CA: Sociology Press.

Glaser, B.G., 1999. The future of grounded theory. *Qualitative Health Research*, 9(6), pp.836-845.

Glaser, B.G., 2002. Constructivist grounded theory? *Forum: qualitative social research*, 3(3.12). [online] Available at: <http://www.qualitative-research.net/index.php/fqs/article/view/825/1792> [Accessed on 22 October 2013].

Glaser, B.G., and Strauss, A.L., 1967. *The discovery of grounded theory*. 5th ed. 2010. London: Aldine Transaction.

Goldstein, H., 2015. Validity, science and educational measurement. *Assessment in Education: Principles, Policy, and Practice*, 22(2), pp.193-201.

Good, T.L. and Brophy, J.E., 2000. *Looking in classrooms*. 8th ed. New York: Longman.

Graue, M.E., 1993. Integrating theory and practice through instructional assessment. *Educational Assessment,* 1(4), pp.283-309.

Greckhamer, T. and Koro-Ljungberg, M., 2005. The erosion of a method: examples from grounded theory. *International Journal of Qualitative Studies in Education*, 18(6), pp.729-750.

Green, J. and Thorogood, N., 2004. *Qualitative methods for health research*. Thousand Oaks, CA: Sage.

Greene, M., 1994. Postmodernism and the crisis of representation. *English Education*, 26(4), pp.206-219.

Grinyer, A., 2002. The anonymity of research participants: assumptions, ethics and practicalities. *Social Research Update*, 36, pp.1-6.

Gronlund, N.E., 1998. *Assessment of student achievement*. 6th ed. Needham Heights, MA: Longwood Division, Allyn and Bacon.

Gronlund, N.E. and Linn, R.L., 1990. *Measurement and evaluation in teaching*. 6th ed. Macmillian, New York.

Guba, E.G., 1990. *The paradigm dialog*. Newbury Park, CA: Sage.

Guba, E.G., 1996. What happened to me on the road to Damascus. In: L. Heshusius and K. Ballard, eds. 1996. *From positivism to interpretivism and beyond: tales of transformation in education and social research*. New York: Teachers College Press, pp.43-49.

Guba, E.G. and Lincoln, Y.S., 1994. Competing paradigms in qualitative research. In: N.K. Denzin and Y.S. Lincoln, eds. 1994. *The Sage handbook of qualitative research*. Thousand Oaks, CA: Sage, pp.105-117.

Guba, E.G. and Lincoln, Y.S., 2005. Pragmatic controversies, contradictions, and emerging confluences. In: N.K. Denzin and Y.S. Lincoln, eds. 2005. *The Sage handbook of qualitative research*. 3rd ed. New Delhi: Sage, pp.191-216.

Gubrium, J.F. and Holstein, J.A., 1997. *The new language of qualitative method*. New York: Oxford University Press.

Gubrium, J.F. and Holstein, J.A., 2001. From the individual interview to the interview society. In: J.F. Gubrium and J.A. Holstein, eds. 2001. *Handbook of interview research: context and methods*. Thousand Oaks: Sage, pp.3-32.

Gullickson, A.R., 1993. Matching measurement instruction to classroom-based evaluation: Perceived discrepancies, needs, and challenges. In: S.T. Wise, ed. 1993. *Teacher training in measurement and assessment skills.* Lincoln, NB: Buros Institute of Mental Measurement, pp.1-25.

Guskey, T.R., 1994. Making the grade: what benefits students? *Educational Leadership*, 52(2), pp.15-20.

Guskey, T.R. 2002. Computerized gradebooks and the myth of objectivity. *Phi Delta Kappan*, 83(10), pp.775-780.

Guskey, T.R., 2004. "Zero" alternatives. *Curriculum and Instructor*. November.

Guskey, T.R., 2006. "It wasn't fair!" Educators' recollections of their experiences as students with grading. In: *Annual meeting of the American educational research association*, April 2006. San Francisco, CA.

Guskey, T.R., 2013. The case against percentage grades. *Educational Leadership*, 71(1), pp.68-72.

Guskey, T., ed., 1996. *Communicating student learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Guskey, T.R. and Bailey, J.M., 2001. *Developing grading and reporting systems for student learning*. Thousand Oaks, CA: Corwin.

Guskey, T. R., and Bailey, J. M., 2010. *Developing standards-based report cards*. Thousand Oaks, CA: Sage.

Hall, P.M. and McGuinty, P.J.W., 2002. Social organization across space and time: the policy process, mesodomain analysis, and breadth of perspective. In: S.C. Chew and D. Knottnerus, eds. 2002. *Structure, culture and history: recent issues in social theory*. Lanham, MD: Rowman and Littlefield, pp.303-322.

Hallberg, L.R.M., 2006. The 'core category' of grounded theory: making constant comparisons. *International Journal of Qualitative Studies on Health and Well-being*, 1, pp.141-148.

Hammersley, M., 2012. Transcription of speech. In: S. Delemont, ed. 2012. *Handbook of qualitative research in education*, pp.439-445.

Hammond, K.R. 1996. *Human judgment and social policy: irreducible uncertainty, inevitable error, unavailable injustice*. New York: Oxford University Press.

Hargreaves, A. 2004. Culture, contracts and change. Keynote address, *Canadian Teachers Federation Conference*. Ottawa.

Harlen, W., 1994. Developing public understanding of education – a role for researchers. *British Educational Research Journal*, 20(1), pp.3-16.

Harlen, W., 2005a. Teachers' summative practices and assessment for learning – tensions and synergies. *The Curriculum Journal*, 16(2), pp.207-223.

Harlen, W., 2005b. Trusting teacher's judgment: research evidence of the reliability and validity of teacher's assessment used for summative purposes. *Research Papers in Education*, 20(3), pp.245-270.

Harlen, W., 2006. On the relationship between assessment for formative and summative purposes. In: J. Gardner, ed., *Assessment and learning*. Los Angeles, CA: Sage, pp.103-118.

Harlen, W., 2012a. On the relationship between assessment for formative and summative purposes. In: J. Gardner, ed., 2012. *Assessment and learning*. 2nd ed. Singapore: Sage, pp.87-102.

Harlen, W., 2012b. The role of assessment in developing motivation for learning. In: J. Gardner, ed., 2012. *Assessment and learning*. 2nd ed. Singapore: Sage, pp.171-184.

Harlen, W. and Deakin Crick, R., 2003. Testing and motivation for learning. *Assessment in Education,* 10(2), pp.169–207.

Harlen, W. and James, M., 1997. Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy and Practice*, 4(3), pp.365-379.

Harrington, A. ed., 2005. *Modern social theory: an introduction*. Kuala Lumpur: Oxford University Press.

Harris, L.R. and Brown, G.T.L., 2009. The complexity of teachers' conceptions of assessment: tensions between the needs of schools and students. *Assessment in Education: Principles, Policy and Practice*, 16(3), pp.365-381.

Hawe, E., 2003. 'It's pretty difficult to fail': the reluctance of lecturers to award a failing grade. *Assessment & Evaluation in Higher Education*, 28(4), pp.371-382.

Hay, P.J. and MacDonald, D., 2008. (Mis)appropriations of criteria and standards-referenced assessment in a performance-based subject. *Assessment in Education: Principles, Policy and Practice*, 15(2), pp.153-168.

Hayward, L., 2015. Assessment is learning: the preposition vanishes. *Assessment in Education: Principles, Policy, and Practice*, 22(2), pp.27-43.

Heath, H. and Cowley, S., 2004. Developing a grounded theory approach: a comparison of Glaser and Strauss. *International Journal of Nursing Studies*, 41(2), pp.141-50.

Herbst, S. and Davies, A., 2014. *A fresh look at grading and reporting*. Bloomington, IN: Solutions Tree Press.

Hesse, M., 1980. *Revolution and reconstruction in the philosophy of science*. Bloomington: Indiana University Press.

Hesse-Biber, S.N., 2007. Teaching grounded theory. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.311-338.

Hoare, K., Mills, J. and Francis, K., 2012. Dancing with data: an example of acquiring theoretical sensitivity in a grounded theory study. *International Journal of Nursing Practice*, 18, pp.240-245.

Hodgson, S., 2004. Cutting through the silence: a sociological construction of self-injury. *Sociological Inquiry*, 74(2), pp.162-179.

Hoge, R.D. and Butcher, R., 1984. Analysis of teacher judgements of pupil achievement levels. *Journal of Educational Psychology*, 76(5), pp.777-781.

Hoge, R.D. and Coladarci T., 1989. Teacher-based judgments of academic achievement: a review of literature. *Review of Educational Research*, 59(3), pp.297-313.

Holstein, J.A. and Gubrium, J.F., 2003. Inside interviewing: new lenses, new concerns. In: J.A. Holstein, and J.F. Gubrium, eds. 2003. *Inside interviewing: new lenses, new concerns.* Thousand Oaks, California: Sage Publications, Inc., pp.3-32.

Holstein, J.A. and Gubrium, J.F., 2011. The constructionist analytics of interpretive practice. In: N.K. Denzin and Y.S. Lincoln, eds., 2011. *The Sage handbook of qualitative research*. 4th ed. Singapore: Sage, pp.341-357.

Holton, J.A., 2007. The coding process and its challenges. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp. 265-289.

Hood, J.C., 2007. Orthodoxy vs. power: the defining traits of grounded theory. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.151-164.

Howley, A., Kusimo, P.S. and Parrot, L., 2001. Grading and the ethos of effort. *Learning Environments Research*, 3, pp.229-246.

Hutchinson, C. and Hayward, L., 2005. The journey so far: assessment for learning in Scotland. *The Curriculum Journal*, 16(2), pp.225-248.

Ingersoll, R.M., 1999. The problem of under-qualified teachers in American secondary schools. *Educational Researcher*, 28(2), pp.26-37.

James, M., 1998. *Using assessment for school improvement*. Oxford: Heinemann Educational.

James, N. and Busher, H., 2009. *Online Interviewing*. Singapore: Sage.

Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation, University of Alberta, 1993. *Principles for fair student assessment practices for education in Canada*. Edmonton: Queen's Printer.

Jootun, D., McGhee, G. and Marland, G.R., 2009. Reflexivity: promoting rigour in qualitative research. *Nursing Standard*, 23(23), pp.42-46.

Josselson, R., 1995. Imagining the real. *Interpreting experience: the narrative study of lives*. Thousand Oaks, CA: Sage.

Kane, M., 1982. A sampling model for validity. *Applied Psychological Measurement,* 6, pp.25-160.

Kane, M. 2001. Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), pp.319-342.

Kelle, U., 2007. The development of categories: different approaches in grounded theory. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.191-213.

Kennedy, B.L., 2009. Infusing participants' voices into grounded theory research: a poetic anthology. *Qualitative Inquiry*, 15(8), pp.1416-1433.

Kennedy, J.T. and Lingard, L.A., 2006. Making sense of grounded theory in medical education. *Medical Education*, 40(2), pp.101-108.

Kincheloe, J.L. and McLaren, P., 2007. Rethinking critical theory and qualitative research. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. 3rd ed. Singapore: Sage, pp.303-342.

Kincheloe, J.L., McLaren, P. and Steinberg, S.R., 2011. Critical pedagogy and qualitative research: moving to the bricolage. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. 4th ed. Singapore: Sage, pp.163-178.

Klapp, A., 2015. Does grading affect educational achievement? A longitudinal study. *Assessment in Education: Principles, Policy, and Practice*, 22(3), pp.302-323.

Klenowski, V. and Wyatt-Smith, C., 2013. *Assessment for education: standards, judgement and moderation*. Sage: London.

Kohn, A. 1993. *Punish by rewards: the trouble with gold stars, incentive plans, A's, high praise, and other bribes*. Boston: Houghton Mufflin.

Kohn, A., 1996. *Beyond discipline: from compliance to community.* Alexandria, VA: Association for Supervision and Curriculum Development.

Koretz, D., 2008. *Measuring up: what educational testing really tells us*. London: Harvard University Press.

Kvale, S., 1994. Ten standard objections to qualitative research interviews. *Journal of Phenomenological Psychology*, 25(20), pp.147-173.

Kvale, S., 1996. *InterViews: an introduction to qualitative research interviewing.* London: Sage.

Kvale, S. and Brinkmann, S., 2009. *InterViews: learning the craft to qualitative research interviewing.* 2nd ed. London: Sage.

Lambert, L.B., 2007. Asking questions of the data: memo writing in the grounded theory tradition. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.245-264.

Larson, M.A., 2009. Stressful, hectic, daunting: a critical policy study of the Ontario teacher performance appraisal system. *Canadian Journal of Educational Administration and Policy*, 95, pp.1-44.

Levin, B., 2004. Media-government relations in education. *Journal of Education Policy*, 19(3), pp.271-283.

Lincoln, Y.S. and Guba, E.G., 1985. *Naturalistic Inquiry*. Newbury Park, CA: Sage Publications.

Lincoln, Y.S., Lynham, S.A. and Guba, E.G., 2011. Paradigmatic controversies, contradictions, and emerging confluences, revisited. In: N.K. Denzin and Y.S. Lincoln, eds., 2011. *The Sage handbook of qualitative research*. 4th ed. Singapore: Sage, pp.97-128.

Linn, R.L., 1994. Performance assessment: policy promises and technical measurements standards. *Educational Researcher*, 23, pp.4-14.

Linn, R.L. and Gronlund, N.E., 1995. *Measurement and assessment in teaching*. New Jersey: Prentice-Hall.

Locke, K., 1996. Rewriting the discovery of grounded theory after 25 years? *Journal of Management Inquiry*, 5(3), pp.239-245.

Lofland, D., Snow, L.A. and Lofland, L.H., 2006. *Analyzing social settings: a guide to qualitative observations and analysis*. 4th ed. Belmond, CA: Wadsworth Thomson.

Lomborg, K. and Kirkevold, M., 2003. Truth and validity in grounded theory – a reconsidered realist interpretation of the criteria: fit, work, relevance and modifiability. *Nursing Philosophy*, 3(4), pp.189-200.

Luke, A. 2003. After the marketplace: evidence, social science and educational research. *Australian Educational Researcher*, 30(2), pp.87-107.

MacDonald, M. and Schreiber, R.E., 2001. Constructing and deconstructing: grounded theory in the postmodern world. In: S.R. Schreiber and P.N. Stern, eds. 2001. *Using Grounded Theory in Nursing*. New York: Springer Publishing Company, pp.177-190.

MacLeans.ca, 2011. McGuinty boats rising graduation rates. [online] Available at <http://oncampus.macleans.ca/education/2011/03/09/mcguinty-boasts-rising-graduation-rates/> [Accessed 10 July 2011].

Malone, L., De Lucchi, L. and Long, K., 2004. All things in moderation. *Science and Children*, 41(5), pp.30-4.

Markham, A.N., 2005. The methods, politics, and ethics of representation in online ethnography. In: N.K. Denzin and Y.S. Lincoln, eds. 2005. *The Sage handbook of qualitative research*. 3rd ed. New Delhi: Sage, pp.793-820.

Martin, N., 2010a. A zero policy in Ontario. *The Winnipeg Free Press*, [online] 26 August 2010. Available at: <http://www.winnipegfreepress.com/opinion/blogs/martin/A-zero-policy-in-Ontario-101564158.html> [Accessed on 2 September 2010].

Martin, N., 2010b. Survey fails no-fail policy. *The Winnipeg Free Press*, [online] 14 January 2010. Available at: <http://www.winnipegfreepress.com/local/survey-fails-no-fail-policy-81449367.html> [Accessed on 2 September 2010].

Marzano, R.J., 2000. *Transforming classroom grading*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Mason, J., 2010. *Qualitative Researching.* 2nd ed. Singapore: Sage.

Matanin, M. and Tannehill, D., 1994. Assessment and grading in physical education. *Journal of Teaching Physical Education*, 13(4), pp.395-405.

Matteucci, M., Tomasetto, C., Patrizia, S. and Carugati F., 2008. Teacher judgments and pupils' casual explanations: social valorization and effort-based explanations in school context. *European Journal of Psychology of Education*, 23(4), pp.421-432.

McGhee, G., Marland, G., and Atkinson, J., 2007. Grounded theory research: literature reviewing and reflexivity. *Journal of Advanced Nursing*, 60(3), pp.334-342.

McMahon, S. and Jones, I., 2015. A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy, and Practice*, 22(3), pp.368-389.

McMillan, J.H., 2001. Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, 20(1), pp.20-32.

McMillan, J.H., 2003. Understanding and improving teachers' classroom assessment decision making: implications for theory and practice. *Educational Measurement: Issues and Practice*, 22(4), pp.34-43.

McMillan, J. H., 2008. *Assessment essentials for standards-based education*. 2nd ed. Thousand Oaks: Corwin Press.

McMillan, J.H., Hellsten, L.M. and Klinger, D.A. 2011. *Classroom assessment: principles and practices for effective standards-based instruction*. Toronto: Pearson Canada.

McMillan, J.H. and Nash, S., 2000. Teacher classroom assessment and grading decision making. *Annual Meeting of the National Council on Measurement in Education*. New Orleans, 27 April 2000. 25-27 April 2000. New Orleans.

McMillan, J.H., Myran, S. and Workman, D., 2002. Elementary teachers' classroom assessment and grading practices. *Journal of Educational Research,* 95(4), pp.203-214.

Mehan, H., 1998. The study of social interaction in educational settings: accomplishments and unresolved issues. *Human Development*, 41(4), pp.245-269.

Meho, L.I., 2005. Email interviewing in qualitative research: a methodological discussion. *Journal of the American Society for Information Science and Technology*, 57(10), pp.1284-1295.

Melia, K.M., 1996. Rediscovering Glaser. *Qualitative Health Research*, 6(3), pp.368-373.

Messick, S., 1989a. Validity. In: R.L. Linn, ed. 1989. *Educational Measurement*. 3rd ed. Phoenix, AZ: Oryx Press, pp.13-103.

Messick, S., 1989b. Meanings and values in test validation: the science and ethics of assessment. *Educational Researcher*, 18(2), pp.5-11.

Messick, S, 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, pp.13-23.

Messick, S, 1995. Standards of validity and validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), pp.5-8.

Miles, M.B. and Huberman, A.M., 1994. *Qualitative data analysis: an expanded sourcebook*. 2nd ed. New Delhi: Sage.

Miles, M.B., Huberman, A.M., and Saldaňa, J., 2014. *Qualitative data analysis: an expanded sourcebook*. 3rd ed. London: Sage.

Millar, E., 2014. Why some schools are giving letter grades a fail. *The Globe and Mail*. [online] 4 April. Available at: <http://www.theglobeandmail.com/news/national/education/schools-that-give-letter-grades-a-fail/article17807841/> Accessed 5 April 2014.

Miller, M.D. and Linn, R.L., 2000. Validation of performance-based assessments. *Applied Psychological Measurement*, 24(4), pp.367-378.

Miller, S.I. and Fredericks, M., 1999. How does grounded theory explain? *Qualitative Health Research*, 9(4), pp. 538-551.

Milliken, P.J. and Schreiber, R., 2012. Examining the nexus between grounded theory and symbolic interactionism. *International Journal of Qualitative Methods*, 11(5), pp.684-696.

Mills, J., Bonnor, A. and Francis, K., 2006a. Adopting a constructivist approach to grounded theory: implications for research design. *International Journal of Nursing Practice*, 12(1), pp. 8-13.

Mills, J., Bonnor, A. and Francis, K., 2006b. The development of constructivist grounded theory. *International Journal of Qualitative Methods*, 5(1), pp.25-35.

Mills, J., Chapman, Y., Bonnor, A. and Francis, K., 2007. Grounded theory: a methodological spiral from positivism to postmodernism. *Journal of Advanced Nursing*, 58(1), pp.72-79.

Mishler, E.G., 1990. Validation in inquiry-guided research: the role of exemplars in narrative studies. *Harvard Educational Review*, 60(4), pp.415-440.

Mislevy, R.J., 1993. Foundations of a new test theory. In: N. Frederiksen, R.J. Mislevy and I.I. Bejay, eds. 1993. *Test theory for a new generation*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Montgomery, P. and Bailey, P.H., 2007. Field notes and theoretical memos in grounded theory. *Western Journal of Nursing Research*, 29(1), pp.65-79.

Morrison, H.G. and Wylie, E.C., 1999. Why national curriculum testing is founded on a methodological thought disorder. *Evaluation and Research in Education*, 13(2), pp.92-105.

Morse, J.M., 1994. Designing funded qualitative research. In: N.K. Denzin and Y.S. Lincon, eds. 1994. *The Sage handbook of qualitative research*. Thousand Oaks, CA: Sage, pp138-157.

Morse, J.M., 1995. The significance of saturation. *Qualitative Health Research*, 5(2), pp.147-149.

Morse, J.M., 2001. Situating grounded theory within qualitative inquiry. In: R. Schreiber and P.N. Stern, eds. *Using Grounded Theory In Nursing*. New York: Springer, pp.1-16.

Morse, J.M., 2006. The politics and evidence. *Qualitative Health Research*, 16(3), pp.395-404.

Morse, J.M., 2007. Sampling in grounded theory. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.229-244.

Morse, J.M., 2009. Tussles, tensions, and resolutions. In: J.M. Morse, P. Noerager Stern, J. Corbin, B. Bowers, K. Charmaz and A.E. Clarke, eds. *Developing grounded theory: the second generation*. 2009. Walnut Creek, CA: Left Coast Press, pp.13-22.

Morse, J.M., Barnett, N., Mayan, M., Olson, K. and Spires, J., 2002. Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods*, 1(2), pp.13-22.

Moss, P.A., 1992. Shifting conceptions of validity in educational assessment: implications for performance assessment. *Review of Educational Research*, 62(3), pp.229-258.

Moss, P.A., 1994. Can there be validity without reliability? *Educational Researcher*, 23(2), pp.5-12.

Moss, P.A., 1995. Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, 14(2), pp.5-13.

Moss, P.A., 2003. Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, 22(4), pp.13-25.

Moss, P.A., 2012. Exploring the macro-micro dynamic in data use practice. *American Journal of Education*, 118(2), pp.223–232.

Mourshed, M., Chijioke, C., and Barber, M., 2010. *How the world's most improved school systems keep getting better*. London: McKinsey and Company.

Mruck, K. and Günter M., 2007. Grounded theory and reflexivity. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.515-538.

Mueller, P.A. and Oppenheimer, D.M., 2014. The pen is mightier than the keyboard: advantages of longhand over laptop note taking. *Psychological Science*, 25(6), pp.1159-1168.

Musial, D., Nieminen G., Thomas J. and Burke K., 2009. *Foundations of meaningful educational assessment*. Kuala Lumpur: McGraw Hill.

Muskin, J.A., 2015. From good ideas to good practice: putting teachers at the centre of education improvement. *Education Assessment, Evaluation Accountability*, 27, pp.93-102.

Neuman, L.W., 2006. *Social research methods: qualitative and quantitative approaches*. 6th ed. Singapore: Pearson.

Newmann, F.M., 1990. Higher order thinking in teaching and social studies: a rationale for the assessment of classroom thoughtfulness. *Journal of Curriculum Studies*, 22(1), pp.41-56.

Newton, P.E., 2003. The defensibility of national curriculum assessment in England. *Research Papers in Education*, 18(2), pp.101-127.

Newton, P.E., 2005. The public understanding of measurement inaccuracy. *British Educational Research Journal*, 31(4), pp.419-442.

Newton, P.E., 2007. Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy and Practice*, 14(2), pp.149-170.

Newton, P.E., 2012. Validity, purpose and the recycling of results from educational assessments. In: J. Gardner, ed. 2012. *Assessment and learning*. 2nd ed. Singapore: Sage, pp.264-276.

Newton, P. and Meadows, M., 2011. Marking quality within test and examination systems. *Assessment in Education: Principles, Policy and Practice*, 18(3), pp.213-216.

Nitko, A.J., 2001. *Educational assessment of students*. New Jersey: Prentice Hall.

OECD (Organisation for Economic Co-operation and Development), 2003. *Key competencies for a successful life and a well-functioning society. The definition and selection of key competencies: executive summary*. S.D. Rychen and L.H. Salganik, eds. Göttingen: Hogrefe and Huber Publishers.

O'Connor, K., 2000. Grading: an exercise in professional judgment. *Orbit*, 30(4), pp.40-42.

O'Connor, K., 2002. *How to grade for learning: linking grades to standards*. Glenview, IL: LessonLab.

O'Connor, K., 2007. *A repair kit for grading: 15 fixes for broken grades*. Singapore: Pearson.

O'Connor, K., 2009. *How to grade for learning: K-12*. 3rd ed. Singapore: Corwin.

O'Connor, K., kenoc@aol.com, 2010. *Researching assessment and evaluation*. [email] Message to G Cavanaugh (kabx9gwc@nottingham.edu.my). Sent 4 March 2010 12:00. [Accessed 5 March 2010].

O'Connor, K., kenoc@aol.com, 2014. *Growing Success*. [email] Message to G Cavanaugh (gordon.william.cavanaugh@gmail.com). Sent 2 October 09:12. [Accessed 2 October 2014].

O'Connor, K. and Wormeli, R., 2011. Reporting student learning. *Educational Leadership*, November, pp.41-44.

O'Donoghue, T., 2007. *Planning your qualitative research project: an introduction to interpretivist research in education*. London: Routledge.

O'Donovan, B., Price, M. and Rust, C., 2004. Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education*, 9(3), pp.325-335.

Office of the Premier, 2015. More Ontario students graduating high school than ever before. *Newsroom.* [online] 1 April. Available at <http://news.ontario.ca/opo/en/2015/04/more-ontario-students-graduating-high-school-than-ever-before.html> [Accessed 15 April 2015].

Oliver, D.G., Serovich, J.M. and Mason, T.L., 2005. Constraints and opportunities with interview transcription: towards reflection in qualitative research. *Social Forces*, 84(2), pp.1273-1289.

Ontario: Ministry of Education, 1999a. *Guide to provincial report cards: grades 9-12*. Toronto: Queen's Printer.

Ontario: Ministry of Education, 1999b. *Ontario secondary schools, grades 9 to 12: program and diploma requirements*. Toronto: Queen's Printer.

Ontario: Ministry of Education, 2000. *The Ontario curriculum, grades 9 to 12: program planning and assessment*. Toronto: Queen's Printer.

Ontario: Ministry of Education, 2008. *Growing Success: Assessment, evaluation and reporting: improving student learning (draft for consultation purposes)*. Toronto: Queen's Printer.

Ontario: Ministry of Education, 2010a. *Growing success: assessment, evaluation, and reporting in Ontario schools*. 1st edition: Covering Grades 1 to 12. Toronto: Queen's Printer.

Ontario: Ministry of Education, 2010b. *New teacher induction program: induction essentials manual*. Toronto: Queen's Printer.

Ontario: Ministry of Education, 2010c. *Student success / learning to 18*. [online] Available at:
<http://www.edu.gov.on.ca/eng/teachers/studentsuccess/strategy.html>> [Accessed 15 September 2010].

Ontario: Ministry of Education, 2011. *Ontario schools: kindergarten to grade 12 policy and program requirements*. Toronto: Queen's Printer.

Ontario: Ministry of Education, 2013a. *Getting results: Ontario's graduation rates*. [online] Available at:
<http://www.edu.gov.on.ca/eng/gettingResultsGrad.html> [Accessed 4 February 2014].

Ontario: Ministry of Education, 2013b. *School effectiveness framework: a support for school improvement and student success*. Toronto: Queen's Printer.

Ontario: Ministry of Education, 2013c. *The Ontario student transcript (OST) manual*. Toronto: Queen's Printer.

Ontario: Ministry of Education, 2014. *Educational facts*. [online] Available at:
<http://www.edu.gov.on.ca/eng/educationFacts.html> [Accessed 28 August 2014].

Ontario: Ministry of Education, 2015. *The Ontario curriculum grades 11 and 12: Canadian and world studies*. Toronto: Queen's Printer.

Ontario: Royal Commission on Learning, 1994. *For the love of learning: report of the royal commission on learning*. Toronto: Queen's Press.

Opie, C., 2004a. Research approaches. In: C. Opie, ed. 2004. *Doing educational research: a guide to first time researchers*. London: Sage, pp.73-93.

Opie, C., 2004b. Research procedures. In: C. Opie, ed. 2004. *Doing educational research: a guide to first time researchers*. London: Sage, pp.94-129.

Outhwaite, W., 2005. Interpretivism and Interactionism. In: Harrington, A., ed. 2005. *Modern social theory: an introduction.* Oxford: Oxford University Press, pp.110-131.

Patton, M.Q., 1999. Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, 34(5), pp.1189-1208.

Patton, M.Q., 2001. *Qualitative Evaluation and Research Methods*. 2ⁿᵈ ed. Thousand Oaks, CA: Sage Publications.

Payne, S., 2007. Grounded theory. In: E. Lyons and A. Coyle, eds. 2007. *Analysing qualitative data in psychology.* London: Sage Publications, pp.65-86.

Pederson, C.H., 2008. Anchors of meaning – helpers of dialogue: the use of images in production of relations and meaning. *International Journal of Qualitative Studies in Education*, 21(1), pp.35–47.

Peel District School Board, 2011. *Peel's Growing Success monograph series: evidence of learning: observations, grades 1-12*. [online] Available at: <http://assessment4learning.com/pdf/PeelSchoolDistrict/Evidence%20of%20Learning%20-%20Observations%201-12%20PDF.pdf> [Accessed 19 March 2015].

Piantanida, M., Tananis, C.S. and Grubs, R.E., 2004. Generating grounded theory of/for educational practice: the journey of three epistemorphs. *International Journal of Qualitative Studies in Education*, 17(3), pp.325-346.

Pidgeon, N. and Henwood, K., 2010. Grounded theory. In: M. Hardy and A. Bryman, eds. 2010. *The Sage handbook of data analysis*. Singapore: Sage, pp.625-648.

Plake, B.S. and Impara, J.C., 1993. Assessment competencies of teachers: a national survey. *Educational Measurement: Issues and Practice*, 12, pp.10-15.

Plake, B. S. and Impara, J.C., 1997. Teacher assessment literacy: what do teachers know about assessment? In: G. D. Phye, ed. *Handbook of classroom assessment: learning, adjustment, and achievement*. 1997. New York: Academic Press, pp.55-68.

Polkinghorne, D. E., 2004. *Practice and the human sciences*. Albany: SUNY Press.

Pollio, H.R. and Humphreys, W.L., 1988. Grading students. In: J.H. McMillan, ed. *Assessing student learning*. 1988. San Francisco: Jossey-Bass.

Pollitt, A., 2012. The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice*, 19(3), pp.281-300.

Popham, W.J., 2003. Seeking redemption for our psychometric sins. *Educational Measurement: Issues and Practice*, 22(1), pp.45-48.

Popham, W.J., 2004. All about accountability/why assessment illiteracy is professional suicide. *Educational Leadership*, 79(2), pp.82-83.

Popham, W.J., 2009. Assessment literacy for teachers: faddish or fundamental? *Theory Into Practice*, 48, pp.4-11.

Preissle, J., 2006. Envisioning qualitative inquiry: a view across four decades. *International Journal of Qualitative Studies in Education*, 19(6), pp.685-695.

Pringle, J., 2007. *John Dewey: a philosopher of education for our time?* London: Continuum.

Pugsley, L., 2012. Walk this way, talk this way: qualitative research on professional education. In: S. Delemont, ed. 2012. *Handbook of Qualitative Research in Education*, pp.157-169.

Punch, K., 2009. *Introduction to research methods in education*. Sage: London.

Raskin, J.D., 2002. Constructivism in psychology: personal construct psychology, radical constructivism, and social constructionism. *American Communication Journal*, 5(3), Spring. [online] Available at: <https://ac-journal.org/journal/vol5/iss3/special/raskin.pdf> [Accessed 21 October 2014].

Ravich, D., 2010. *The death and life of the great American school system: how testing and choice are undermining education*. New York: Basic Books.

Reeves, D.B., 2004. The case against zero. *Phi Delta Kappan*, 86(4), pp.324-325.

Reeves, D.B., 2008. Toxic grading practices. *The Ontario ministry of education assessment and evaluation symposium*. Ottawa, Canada, 8 January 2008. [online] Available at: <http://www.youtube.com/watch?v=jduiAnm-O3w> [Accessed 20 July 2011].

Reeves, D.B., 2010. *Elements of grading*. Bloomington, IN: Solution Tree.

Reichertz, J., 2007. Abduction: The logic of discovery of grounded theory. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.214-228.

Reiman, A. and Johnson, L., 2003. Promoting teacher professional judgment. *Journal of Research in Education*, 13(1), pp.4-14.

Rennie, D.L., 1998. The pressing need for a coherent logic of justification. *Theory & Psychology*, 8(1), pp.101-119.

Rennie, D.L., 2000. Grounded theory methodology as methical hermeneutics: reconciling realism and relativism. *Theory & Psychology*, 10(4), pp.481-502.

Richardson, L. and St. Pierre, E.A., 2005. Writing: a method of inquiry. In: N.K. Denzin and Y.S. Lincon, eds. 2005. *The Sage handbook of qualitative research*. 3rd ed. New Delhi: Sage, pp.959-978.

Robrecht, L., 1995. Grounded theory: evolving methods. *Qualitative Health Research*, 5(2), pp.169-177.

Ross, J.A. and Kostuch, L., 2011. Consistency of report card grades and external assessments in a Canadian province. *Educational Assessment Evaluation and Accountability*, 23(2), pp.159-180.

Rolfe, G., 2006a. Judgements without rules: towards a postmodern ironist concept of research validity. *Nursing Inquiry*, 13(1), pp.7-15.

Rolfe, G, 2006b. Validity, trustworthiness and rigour: quality and the idea of qualitative research. *Journal of Advanced Nursing*, 53(3), pp.304-10.

Rowe, K. J. and Hill, P. W., 1996. Assessing, recording and reporting students' educational progress: the case for 'subject profiles.' *Assessment in Education: Principles, Policy and Practice,* 3(3), pp.309–352.

Rushowy, K., 2015. High school grad rates raise slightly in Ontario. *The Toronto Star*. [online] Available at: <http://www.thestar.com/yourtoronto/education/2015/04/01/high-school-grad-rates-rise-slightly-in-ontario.html> [Accessed 2 April 2015].

Sadler, D.R. 1985. The origins and functions of evaluative criteria. *Educational Theory*, 35(3), pp.285-297.

Sadler, D.R., 1987. Specifying and promulgating achievement standards. *Oxford Review of Education*, 13(2), pp.191-209.

Sadler, D.R., 1998. Formative assessment: revisiting the territory. *Assessment in Education: Principles, Policy and Practice*, 5(1), pp.77–85.

Sadler, D.R. 2009. Indeterminacy in the use of preset criteria for assessment and grading. *Assessment and Evaluation in Higher Education*, 34(2), pp.159–179.

Sandelowski, M., 1993. Rigor or rigor mortis: the problem of rigor in qualitative research revisited. *Advances in Nursing Science*, 16(2), pp.1-8.

Scates, D.E., 1943. Difference between measurement criteria of pure scientists and of classroom teachers. *Journal of Educational Research*, 37, pp.1-13.

Schafer, W.D., 1991. Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice*, 10(1), pp.7-12.

Schafer, W.D. and Lissitz, R.W., 1987. Measurement training for school personnel: recommendations and reality. *Journal of Teacher Education*, 38(3), pp.57-63.

Schunk, D. H., 1996. Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal*, 33(2), pp.359-382.

Schwandt, T.A., 1996. Farewell to criteriology. *Qualitative Inquiry*, 2(1), pp.58-76.

Schwandt, T.A., 1999. On understanding understanding. *Qualitative Inquiry*, 5(4), pp.451-464.

Schwandt, T.A., 2000. Three epistemological stances for qualitative inquiry: interpretivism, hermeneutics, and social constructionism. In: N.K. Denzin and Y.S. Lincoln, eds. *Handbook of qualitative research*. 2nd ed. Thousand Oaks, CA: Sage, pp.189-213.

Scott, S., Webber, C.F., Lupart, J.L., Aitken, N. and Scott, S.E., 2014. Fair and equitable assessment practices for all students. *Assessment in Education: Principles, Policy, and Practice*, 21(1), pp.52-70.

Scriffiny, P., 2008. Seven reasons for standards-based grading. *Educational Leadership*, 66(2), pp.70-74.

Sebeok, T. and Sebeok, J.U., 1981. You know my method: a juxtaposition of Charles S. Peirce and Sherlock Holmes. In: T. Sebeok, ed. 2009. *The play of musement.* Bloomington: Indiana University Press, pp.17-52. [online] Available at: <http://www.visual-memory.co.uk/b_resources/abduction.html> [Accessed 20 November 2014].

Selby, D., and Murphy, S., 1992. Graded or degraded: perceptions of letter-grading for mainstreamed learning-disabled students. *British Columbia Journal of Special Education*, 16(1), pp.92-104.

Seligman, M., 1998. *Learned optimism*. New York: Pocket Books.

Shaw, I., 2003. Ethics in qualitative research and evaluation. *Journal of Social Work*, 3(1), pp. 9-29.

Shepard, L.A., 1989. Why we need better assessments. *Educational Leadership*, 46(7), pp.4-9.

Shepard, L.A., 1993. Evaluating test validity. *Review of Research in Education*, 19, pp.405-450.

Shepard, L.A., 1995. Using assessment to improve learning. *Educational Leadership*, 52(5), pp.38-43.

Shepard, L.A., 1997. *Measuring achievement: what does it mean to test for robust understandings?* Princeton, NJ: Educational Testing Service.

Shepard, L.A., 2000a. The role of assessment in a learning culture. *Educational Researcher*, 29(7), pp.4-14.

Shepard, L.A., 2000b. The role of classroom assessment in teaching and learning. In: V. Richardson, ed. 2000. *Handbook of research on teaching*. 4th ed. Washington, DC: American Educational Research Association, pp.1066-1101.

Shepard, L.A., 2009. Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice*, 28(3), 32–37.

Sikes, P., 2004. Methodology, procedures and ethical concerns. In: C. Opie, ed., *Doing educational research: a guide to first time researchers*. London: Sage, pp.15-33.

Silverman, D., 2005. *Doing qualitative research*. 2nd ed. London: Sage.

Simon, M., Tierney, R.D., Forgette-Giroux, R., Charland, J., Noonan, B., and Duncan, R., 2010. A secondary school teacher's description of the process of determining report card grades. *McGill Journal of Education*, 45(3), pp.535-554.

Simons, H. and Usher, R., eds., 2000. *Situated ethics in educational research*. London: Routledge Falmer.

Smith, J.K., 2003. Reconsidering reliability for classroom assessment and grading. *Educational Measurement: Issues and Practice*, 22(4), pp.26-33.

Smith, J., 2014. Why principals may be the key to your child's success at school. *The Globe and Mail*. [online] 2 September 2014. Available at: <http://www.theglobeandmail.com/globe-debate/why-principals-may-be-the-key-to-your-childs-success-at-school/article20296502/> [Accessed 2 September 2014].

Smyth, M. and Williamson, E., 2004. *Researchers and their "subjects": ethics, power, knowledge and consent*. Bristol: Policy Press.

Speck, B. W., 1998. Unveiling some of the mystery of professional judgment in classroom assessment. In: R. S. Anderson and B. W. Speck, eds. 1998. *Changing the way we grade student performance: classroom assessment and the new learning paradigm*. San Francisco, CA: Jossey-Bass, pp.89-96.

Stainton-Rogers, W., 2006. Logics of enquiry. In S. Potter, ed. 2006. *Doing postgraduate research*. 2nd ed. London: Sage, pp. 73-92.

Stake, R. 2004. *Standards-based responsive evaluation.* Thousand Oaks, CA: Sage.

Star, S.B., 2007. Living grounded theory: cognitive and emotional forms of pragmatism. In A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.75-93.

Steffenhagen, J., 2010. Zeros now possible when Ontario students missed deadlines. *The Vancouver Sun*. [online] 29 August. Available at: <http://communities.canada.com/vancouversun/blogs/reportcard/archive/2010/08/29/zeros-possible-when-ontario-students-miss-deadlines.aspx> [Accessed on 1 September 2010].

Stern, P.N., 1980. Grounded theory methodology: its uses and processes. *Images*, 12(1), pp.20-23.

Stern, P.N., 1995a. Eroding grounded theory. In J.M. Morse, ed. *Critical issues in qualitative research methods*. Thousand Oaks, CA: Sage, pp. 212-223.

Stern, P.N., 1995b. Grounded theory methodology: its uses and processes. In B.G. Glaser, ed., *Grounded theory 1984-1994*, volume 1. Mill Valley, CA: Sociology Press.

Stern, P.N., 2007. On solid ground: Essential properties for growing grounded theory. In A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp.114-126.

Stern, P.N., 2009a. Glaserian grounded theory. In J.M. Morse, P. Noerager Stern, J. Corbin, B. Bowers, K. Charmaz and A.E. Clarke, eds. *Developing grounded theory: the second generation*. 2009. Walnut Creek, CA: Left Coast Press, pp.55-85.

Stern, P.N., 2009b. In the beginning Glaser and Strauss invented grounded theory. In J.M. Morse, P. Noerager Stern, J. Corbin, B. Bowers, K. Charmaz and A.E. Clarke, eds. *Developing grounded theory: the second generation*. 2009. Walnut Creek, CA: Left Coast Press, pp.22-33.

Stiggins, R.J., 1988. Revitalizing classroom assessment: the highest instructional priority. *Phi Delta Kappan*, 69(5), pp.363-368.

Stiggins, R.J., 1991. Assessment literacy. *Phi Delta Kappan*, 72, pp.534-539.

Stiggins, R.J., 1992. High quality classroom assessment: what does it really mean? *Educational Measurement: Issues and Practice*, 11(2), pp.35-39.

Stiggins, R. J., 1997. *Student-centered classroom assessment*. 2nd ed. Upper Saddle River, NJ: Prentice-Hall, Inc.

Stiggins, R.J., 1999. Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18(1), pp.23-27.

Stiggins, R.J., 2001a. *Student-involved classroom assessment*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Stiggins, R.J., 2001b. The unfulfilled promise of classroom assessment. *Educational Measurements: Issues and Practice*, 20(3), pp.5-14.

Stiggins, R.J., 2005. *Student-involved assessment for learning*. 4th ed. Upper Saddle River, NJ: Pearson.

Stiggins, R.J. and Bridgeford, N.J., 1985. The ecology of classroom assessment. *Journal of Educational Measurement*, 22(4), pp.271-287.

Stiggins, R.J. and Conklin, N.F., 1992. *In teachers' hands: investigating the practices of classroom assessment*. Albany: State University of New York Press.

Stiggins, R.J. and Duke, D.L., 1991. District grading policies and their potential impact on at-risk students. In: *Annual meeting of the American Educational Research Association*, April 1991. Chicago, IL.

Stiggins, R. J., Frisbie, D. A., and Griswold, P.A., 1989. Inside high school grading practices: building a research agenda. *Educational Measurement: Issues and Practice,* 8(2), pp.5-14.

Stobart, G., 2005. Fairness in multicultural student assessment systems. *Assessment in Education: Principles, Policy and Practice*, 12(3), pp.275-287.

Stobart, G., 2008. *Testing times: the uses and abuses of assessment*. London: Taylor and Francis.

Strauss, A.L., 1958. *Mirrors and masks: the search for identity*. 1997. Piscataway, NJ: Transaction Publishers.

Strauss, A.L., 1987. *Qualitative analysis for social scientists*. Cambridge: Cambridge University Press.

Strauss, A.L., 1993. *Continual permutations of action*. New York: Adine de Gruyter.

Strauss, A.L. and Corbin J.M., 1990. *The basics of qualitative research: techniques and procedures for developing grounded theory*. 1st ed. Thousand Oaks, CA: Sage.

Strauss, A.L. and Corbin J., 1994. Grounded theory methodology: an overview. In: N.K. Denzin and Y.S. Lincolin, eds. 1994. *The Sage Handbook of Qualitative Research*, 1st ed. Thousand Oaks, CA: Sage, pp.273-285.

Strauss, A.L. and Corbin J., 1996. Analytic ordering for theoretical purposes. *Qualitative Inquiry*, 2(2), pp.139-150.

Strauss, A.L. and Corbin J., 1997. *Grounded theory in practice*. Thousand Oaks, CA: Sage.

Strauss, A.L. and Corbin J., 1998. *The basics of qualitative research: techniques and procedures for developing grounded theory*. 2nd ed. Thousand Oaks, CA: Sage.

Strübing, J., 2007. Research as pragmatic problem-solving: the pragmatist roots of empirically-grounded theorizing. In: A. Bryant and K. Charmaz, eds. 2007. *The Sage handbook of grounded theory*. Singapore: Sage, pp. 580-601.

Suddaby, R., 2006. From the editors: what grounded theory is not. *Academy of Management Journal*, 49(4), pp.633-642.

Sutton, R., 1991. *Assessment: a framework for teachers.* Berkshire: The Nfer-Nelson Publishing Company Ltd.

Taylor, S.J. and Bogdan, R., 1998. *Introduction to qualitative research methods: guidebook and resource*. 3rd ed. Singapore: John Wiley & Sons, Inc.

Terwilliger, J.S., 1989. Classroom standard setting and grading practices. *Educational Measurement: Issues and Practice*, 8(2), pp.15-19.

Thompson, C., 1999. Qualitative research into nurse decision making: factors for consideration in theoretical sampling. *Qualitative Health Research,* 9(6), pp.815-828.

Thornberg, R., 2007. Inconsistencies in everyday patterns of school rules. *Ethnography and Education*, 2(3), pp.401-416.

Thornberg, R., 2009. The moral construction of the good pupil embedded in school rules. *Education, Citizenship and Social Justice*, 4(3), pp.245-261.

Tierney, J., 2012. Why do so many teachers quit their jobs? Because they hate their bosses. *The Atlantic* [online] 16 November. Available at: <http://www.theatlantic.com/national/archive/2012/11/why-do-so-many-teachers-quit-their-jobs-because-they-hate-their-bosses/265310/> [Accessed 2 Oct 2013].

Tittle, C.K., 1989. Validity: whose construction is it in the testing and learning context? *Educational Researcher,* 8(1), pp.5-13, 34.

Tittle, C.K., 1994. Toward an educational psychology of assessment for teaching and learning: theories, contexts, and validation arguments. *Educational Psychologist*, 29, pp.149-162.

Toma, J.D., 2000. How getting close to your subjects makes qualitative research better. *Theory Into Practice*, 39(3), pp.177-184.

Troug, A.L. and Friedman, S.J., 1996. Evaluating high school teachers' written grading polices. *Annual meeting of the national council on measurement in education*. New York, 9-11 April 1996. New York.

Trouilloud, D.O., Sarrazin, P.G., Martinek, T.J., and Guillet, E., 2002. The influence of teacher expectations on student achievement in physical education classes: Pygmalion revisited. *European Journal of Social Psychology*, 32(5), pp.591-607.

University of Nottingham, 2010. *University of Nottingham Code of Research Conduct and Research Ethics*. 3rd ed. Nottingham: University of Nottingham Press.

Urquhart, C., 2002. Re-grounding grounded theory or reinforcing old prejudices? A brief response to Bryant. *Journal of Information Technology and Application*, 4(3), pp.43-54.

Volante, L., 2006. Reducing bias in classroom assessment and evaluation. *Orbit,* 36(2), pp.34-36.

Volante, L., 2007. Educational quality and accountability in Ontario: past, present, and future. *Canadian Journal of Educational Administration and Policy*, 58, pp.1-21.

Walvoord, BE. and Anderson, V.J., 1998. *Effective grading: a tool for learning and assessment*. San Francisco: Jossey-Bass Publishers.

Warren, C., 2001. Qualitative interviewing. In: J. Gubrium and J. Holstein, eds. 2001. *Handbook of interview research: context & method*. Thousand Oaks, California: Sage, pp.83-102.

Wasserman, J.A., Clair, J.M. and Wilson, K.L., 2009. Problematics of grounded theory: innovations for developing an increasingly rigorous qualitative method. *Qualitative Research*, 9(3), pp.355-381.

Wenger, E., 1998. *Communities of practice: learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Western and Northern Protocol for Collaboration in Education, 2006. *Rethinking classroom assessment with purpose in mind*. Winnipeg: Manitoba Education, Citizenship and Youth.

Whittington, D., 1999. Making room for values and fairness: teaching reliability and validity in the classroom context. *Educational Measurement: Issues and Practice*, 18(1), pp.14-22, 27.

Wiggins, G., 1993. Assessment, authenticity, context, and validity. *Phi Delta Kappan*, 75(3), pp.200-214.

Wilen, W., Ishler, M., Hutchinson, J., and Kindsvatter, R., 2004. *Dynamics of effective teaching*. 5th ed. New York: Addison Wesley Longman.

Willig, C., 2008. Introducing qualitative research in psychology. Berkshire: McGraw-Hill Education.

Wilson, H.S. and Hutchinson, S.A., 1996. Methodological mistakes in grounded theory. *Nursing Research*, 45(2), pp.122-124.

Wilson, N., 1998. Educational standards and the problem of error. *Educational Policy Analysis Archives*, 6(10).

Wolcott, H., 2009. *Writing up qualitative research*. 3rd ed. Los Angeles: Sage.

Wormeli, R. 2006a. *Fair isn't always equal: assessing and grading in the differentiated classroom*. Portland, ME; Stenhouse Publishers.

Wormeli, R., 2006b. Turning zeros to 60s. *Middle Ground: The Magazine for Middle Level Education*, 9(3), pp. 12-15.

Wuest, J., 2001. Precarious ordering: toward a formal theory of women's caring. *Health Care for Women International*, 22, pp.167-193.

Wyatt-Smith, C., 1999. Reading for assessment: how teachers ascribe meaning and value to student writing. *Assessment in Education: Principles, Policy and Practice*, 6(2), pp.195-223.

Wyatt-Smith, C.M. and Castleton, G., 2005. Examining how teachers judge student writing: an Australian case study. *Journal of Curriculum Studies*, 37(2), pp.131-54.

Wyatt-Smith, C.M. and Klenowski, V., 2013. Explicit, latent and meta-criteria: types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy and Practice*, 20(1), pp.35-52.

Wyatt-Smith, C.M., Klenowski, V. and Gunn S., 2010. The centrality of teachers' judgement practices in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy and Practice*, 17(1), pp.59-75.

Yates, L., 2004. *What does good education research look like?* Maidenhead, England: Open University Press.

Young, V.M. and Kim, D.H., 2010. Using assessments for instructional improvement: a literature review. *Educational Policy Analysis Archives,* 18(19).

Zegarac, G. and Franz, R., 2007. Secondary school reform in Ontario and the role of research, evaluation and indicator data. *American Educational Research Association*. Chicago, 11 April 2007. Toronto: Ontario Ministry of Education.

Zeidner, M., 1992. Key facets of classroom grading: A comparison of teacher and student perspectives. *Contemporary Educational Psychology*, 17(3), pp.224–243.

Zhu, M. and Urhahne, U., 2015. Teacher judgements of students' foreign-language achievement. *European Journal of Psychological Education*, 30, pp.21-39.

Zirkel, P.A., 1999. Grade inflation: a leadership opportunity for schools of education? *Teachers College Record*, 101, pp.247-260.

Zwaagstra, M., 2012a. Keep no-zero policies out of schools in Nova Scotia. *The Chronicle Herald*. [online] 3 August. Available at: <http://www.thechronicleherald.ca/opinion/131227-keep-no-zero-policies-out-of-schools-in-nova-scotia> [Accessed 5 August 2012].

Zwaagstra, M., 2012b. The folly of 'no-zero' policies in schools. *The Winnipeg Free Press*. [online] 4 September. Available at: <http://www.winnipegfreepress.com/opinion/westview/the-folly-of-no-zero-policies-in-schools-168428526.html> [Accessed 5 September 2012].

APPENDIX ONE: Abbreviations

CGT -   Constructivist Grounded Theory

GTM -   Grounded Theory Methods

NTIP -   New Teacher Induction Program

OECD -   Organisation for Economic Co-Operation and Development

OME -   Ontario Ministry of Education

SA -   Situational Analysis

APPENDIX TWO: Sample Participant Consent Form

Project Title: Professional Judgement and Assessment of Learning Practices in Ontario English High Schools

Researcher's Name: Gordon William Cavanaugh

Supervisor's Name: Ganakumaran Subramaniam

- I have read the Participant Information Sheet and the nature and purpose of the research project has been explained to me. I understand and agree to take part
- I understand the purpose of the research project and my involvement in it
- I understand that I may withdraw from the research project at any stage and that this will not affect my status now or in the future
- I understand that while information gained during the study may be published, I will not be identified and my personal results will remain confidential
- I understand that I will be audiotaped during the interview
- I understand that data will be stored at the private address of the researcher
- I understand that I may contact the researcher or supervisor if I require further information about the research, and that I may contact the Research Ethics Coordinator of the School of Education, University of Nottingham, if I wish to make a formal complain relating to my involvement in the research

Signed ………………………………………………….. (research participant)

Print Name ……………………………………………….. **Date** ………………………

Contact Details

Researcher: Gordon William Cavanaugh Gordon.cavan@taylors.edu.my

Supervisor: Ganakumaran Subramaniam +60389248691

School of Education Research Ethics Coordinator

educationresearchethics@nottingham.ac.uk